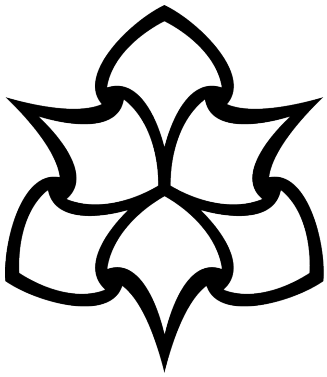


The Development of a Framework for Semantic Similarity Measures for the Arabic Language

Faaza Abduljabar Al-Marsoomi

A thesis submitted in partial fulfilment of the requirements of the Manchester Metropolitan University for the degree of Doctor of Philosophy

School of Computing, Mathematics and Digital Technology
The Manchester Metropolitan University

January 2015

Abstract

This thesis presents a novel framework for developing an Arabic Short Text Semantic Similarity (STSS) measure, namely that of NasTa. STSS measures are developed for short texts of 10 -25 words long. The algorithm calculates the STSS based on Part of Speech (POS), Arabic Word Sense Disambiguation (WSD), semantic nets and corpus statistics.

The proposed framework is founded on word similarity measures. Firstly, a novel Arabic noun similarity measure is created using information sources extracted from a lexical database known as Arabic WordNet. Secondly, a novel verb similarity algorithm is created based on the assumption that words sharing a common root usually have a related meaning which is a central characteristic of Arabic language. Two Arabic word benchmark datasets, noun and verb are created to evaluate them. These are the first of their kinds for Arabic. Their creation methodologies use the best available experimental techniques to create materials and collect human ratings from representative samples of the Arabic speaking population. Experimental evaluation indicates that the Arabic noun and the Arabic verb measures performed well and achieved good correlations comparison with the average human performance on the noun and verb benchmark datasets respectively.

Specific features of the Arabic language are addressed. A new Arabic WSD algorithm is created to address the challenge of ambiguity caused by missing diacritics in the contemporary Arabic writing system. The algorithm disambiguates all words (nouns and verbs) in the Arabic short texts without requiring any manual training data. Moreover, a novel algorithm is presented to identify the similarity score between two words belonging to different POS, either a pair comprising a noun and verb or a verb and noun. This algorithm is developed to perform Arabic WSD based on the concept of noun semantic similarity.

Important benchmark datasets for text similarity are presented: ASTSS-68 and ASTSS-21. Experimental results indicate that the performance of the Arabic STSS algorithm achieved a good correlation comparison with the average human performance on ASTSS-68 which was statistically significant.

Copyright

Declaration

I declare that no portion of the work referred to in the thesis has been previously submitted for another degree or qualification at any other university or other institute of learning.

_____

Faaza Al-Marsoomi

Acknowledgement

All praise is due to God,
the Creator and Sustainer of the Universe.

I would like to express my sincere gratitude to my project supervisors, Dr James O'Shea, Dr Keeley Crockett and Dr Zuhair Bandar, for their guidance, advice and continuous support throughout the production of this research and thesis. I would also like to thank the Arabic linguistic, Dr Khalid Abood for his assistance and advice during my study.

I would like to thank my parents and my sisters, Israa, Asma and Maada for their unwavering support and encouragement throughout the years.

Finally, I would like to thank my husband, Ahmed, and my wonderful sons, Mustafa, Osamah and Aihem for having the patience and understanding to allow me to pursue my dream. I couldn't have done it without you.

Contents

List of Figures

List of Tables

List of Abbreviations

A	Agent
ANSS	Arabic Noun Semantic Similarity
ASTSS	Arabic Short Text Semantic Similarity
AVN	Arabic VerbNet
AVSS	Arabic Verb Semantic Similarity
AWC	Arabic Word Count
AWN	Arabic WordNet
AWSAD	All Word Sense Arabic Disambiguation
B	Benefactive
BAMA	Buckwalter Arabic Morphological Analyser
CG	Case Grammar
C-Structure	Constituent-Structure
E	Experiencer
F-Structure	Functional-Structure
HIT	Human Intelligence Task
HSM	High Similarity of Meaning
IC	Information Content
KalTa-A	تشابه الكلمة – اسم ***Kalimap TashaAboh – Aisom*** "Word Similarity-Noun"
KalTa-AF	تشابه الكلمة–اسم فعـل ***Kalimap TashaAboh – Aisom FiEol*** "Word Similarit- Noun Verb"
KalTa-F	تشابه الكلمة – فعـل ***Kalimap TashaAboh – FiEol*** "Word Similarity-Verb"
L	Locative
LAN	List of Arabic Nouns
LCS	Lowest Common Subsumer
LFG	Lexical Functional Grammar
LOV	List of Original Verbs
LSA	Latent Semantic Analysis
LSM	Low Similarity of Meaning
M&C	Miller & Charles
MRS	Microsoft Research
MSA	Modern Standard Arabic
MSM	Medium Similarity of Meaning

NasTa تشابــه النــص ***Nas TashaAboh*** "Short Text Similarity"

NasTa-A تشابــه النــص - اسم ***Nas TashaAboh– Aisom*** "Short Text Similarity-Noun"

NasTa-F تشابــه النــص - فعل ***Nas TashaAboh– FiEol*** "Short Text Similarity-Verb"

NLP Natural Language Processing

O Object

PARC Palo Alto Research Centre (PARC)

ParGram Parallel Grammar

PATB Penn Arabic Treebank

POS Parts of Speech

R&G Rubinstein and Goodenough

SD Standard Deviation

STSS Short Text Semantic Similarity

SUMO Suggested Upper Merged Ontology

SVO Subject-Verb –Object

synset synonym set

WO Word Order

WSD Word Sense Disambiguation

WSD-STS Word Sense Disambiguation-based Sentence Similarity

XLE Xerox Linguistic Environment

List of Glossary

1. **An Arabic Equational Sentence** is a sentence without a verb and its structure consists of the subject and predicate.
2. **A Category Norm** is defined as a set of words within the same theme, listed by frequency, which is created as responses by human participants to a specific category.
3. **A Verbal Sentence** is a regular sentence in Arabic and its structure consists of the verb, subject and object.
4. **Lemmatisation** is the task of finding the canonical form, or dictionary form, (which is also named the lemma) for words.
5. **Parsing** is the process of assigning a syntactic structure to a group of words and is automatically done using the text parser technique.
6. **POS Tagging** is the process of assigning a word class (grammatical category label) to each word in a text and is automatically performed using the POS tagger technique.
7. **Word Sense Disambiguation** is defined as the process of identifying the correct sense to a particular word based on the context in which it appears.
8. **Word Structure or Morphology** concerns the regulation, rules, and processes of the meaningful units of language, in terms of whether these units are words or parts of words, such as different kinds of affixes.

Chapter 1

Introduction

1.1 Contribution

This thesis presents research investigating a computational approach to Arabic short text semantic similarity, the similarity of the short text meaning. Short Text Semantic Similarity (STSS) measurements are developed to measure the similarity between very short texts of 10 -25 words long. This is the length of typical utterances in human dialogue (O'Shea et al., 2010). The importance of STSS measures is growing due to the large number of applications that are emerging in numerous text-related research fields. For example, in web page retrieval, STSS measures are used for the improvement of the retrieval effectiveness by means of the calculation of the similarities of page titles (Park et al., 2005). Text mining can benefit from the use of STSS measure as criterion to detect concealed knowledge from textual databases (Atkinson-Abutridy et al., 2004). In conversational agent, the employment of STSS measure can greatly reduce the scripting process through the use of natural sentences instead of large numbers of structural patterns containing wildcards (O'Shea, K. et al., 2010).

Unfortunately, research in the semantic similarity field has neglected the Arabic language. (Habash, 2010) reported that the research into Arabic computational semantics is much smaller than the research in other areas in Natural Language Processing (NLP) mainly due to the higher complexity and subtlety of Arabic. Despite this challenge, novel work on STSS using Arabic is presented in this thesis. The main contributions of the work in this thesis fall into three areas:

- The contribution to the automatic measurement of Arabic semantic similarity. This includes a novel framework for developing an Arabic STSS measure which is the most significant contribution of the work in this thesis. Also, two novel Arabic word semantic similarity measures have been created: Arabic noun semantic similarity and Arabic verb semantic similarity. These measures are

expected to contribute to the development of the performance of many Arabic applications.

- The contribution to Arabic semantic similarity resources. The production of two Arabic short text benchmark datasets for evaluating and optimizing the proposed STSS measurement algorithms. Similarly, two Arabic word benchmark datasets for evaluating the Arabic noun semantic similarity algorithms and the Arabic verb similarity algorithms. These datasets are the first of their kinds for Arabic. It is expected that these datasets will be regarded as a reference basis from which to evaluate and compare different methodologies in the field.

- The contribution to Arabic Word Sense Disambiguation (WSD). The development of a new Arabic WSD algorithm to disambiguate all words (nouns and verbs) in the Arabic short texts without requiring any manual training data. Moreover, a novel algorithm is presented to identify the similarity score between two words which have different Parts of Speech (POS), either a pair comprising a noun and a verb or a verb and a noun. This algorithm is developed to perform Arabic WSD based on the concept of noun semantic similarity.

1.2 Research Questions

1. Is it possible to construct a framework for developing a short text semantic similarity measure for Arabic language?
2. Are there features of Arabic language which would prevent the construction of the framework for semantic similarity?
3. Do the necessary components exist for constructing a measure with a framework?
4. Where there are missing components from NLP that are required, is it possible to create these for the Arabic language? i.e.
    - Is it possible to measure the semantic similarity between a pair of Arabic nouns?
    - Is it possible to measure the semantic similarity between a pair of Arabic verbs?

- Is it possible to disambiguate all words in an Arabic short text?
- Is it possible to measure the similarity between Arabic words belonging to a different POS?

5. Is it possible to create suitable benchmark datasets for noun, verb, and STSS algorithms test?

1.3 Hypotheses

1. **H_0 (Null Hypothesis):** it is not possible for a machine based Arabic noun semantic similarity measure to re-produce human intuitive measures of semantic similarity.
   H_1: it is possible for a machine based Arabic noun semantic similarity measure to re-produce human intuitive measures of semantic similarity.

2. **H_0:** it is not possible for a machine based Arabic verb semantic similarity measure to re-produce human intuitive measures of semantic similarity.
   H_1: it is possible for a machine based Arabic verb semantic similarity measure to re-produce human intuitive measures of semantic similarity.

3. **H_0:** it is not possible for an Arabic algorithm for all-word sense disambiguation to achieve the same classification as human would make.
   H_1: it is possible for an Arabic algorithm for all-word sense disambiguation to achieve the same classification as human would make.

4. **H_0:** it is not possible for a machine based Arabic short text semantic similarity measure to re-produce human intuitive measures of semantic similarity.
   H_1: it is possible for a machine based Arabic short text semantic similarity measure to re-produce human intuitive measures of semantic similarity.

5. **H_0**: it is not possible to construct a noun dataset for Arabic within a limited size which effectively represents human intuition.
   H_1: it is possible to construct a noun dataset for Arabic within a limited size which effectively represents human intuition.

6. **H_0**: it is not possible to construct a verb dataset for Arabic within a limited size which effectively represents human intuition.
   H_1: it is possible to construct a verb dataset for Arabic within a limited size which effectively represents human intuition.

7. **H_0:** it is not possible to construct a short text dataset for Arabic within a limited size which effectively represents human intuition.
   H_1: it is possible to construct a short text dataset for Arabic within a limited size which effectively represents human intuition.

1.4 Thesis Outline

The rest of this thesis is organised in three parts. First part comprises of two chapters (2 and 3) which presents a background material to this thesis. The second part (chapter 4) presents the Arabic STSS framework whilst the third part (chapter 5 and 6) concerns with creation of four datasets in order to use them in the evaluation process of the proposed framework's algorithms.

Chapter 2 describes the characteristics of the Arabic language and their influence on the STSS computation. This chapter reviews the Arabic NLP techniques used in text pre-processing. The main source of Arabic ambiguity and the current state of Arabic WSD algorithms created to manage this challenge are also reviewed in this chapter.

Chapter 3 reviews the current state of STSS measures and highlights the major challenges faced by the existing measures. This chapter reviews the current state of word similarity measures which are considered to be the main requirement of the creation of the STSS measure. The current states of word and short text benchmark datasets used in the evaluation processes of the existing word and STSS measures are also reviewed in this chapter with highlighting the challenges of the dataset design process.

Chapter 4 presents a novel framework (NasTa) for developing a measurement algorithm to calculate the semantic similarity between two Arabic short texts. The development process of NasTa consists of two phases. This chapter describes the

NasTa components of each phase with the novel algorithms that has been created to meet its requirements. This includes three Arabic word similarity measures and an Arabic word sense disambiguation algorithm.

Chapter 5 describes the production of the first two Arabic word similarity benchmark datasets and their creation methodologies: the Arabic noun benchmark dataset and the Arabic verb benchmark dataset. These datasets are used to validate of the Arabic word (noun and verb) measures presented in chapter 4. This chapter also describes the evaluation procedure of each measure which involves the creation of the training sub-dataset to use in the parameter optimization process and evaluation sub-datasets to use in the process of validating of the Arabic word measure.

Chapter 6 describes the production of the first Arabic short text benchmark dataset (ASTSS-68) with its creation methodology. The motivation of the creation of this dataset is to evaluate the ASTSS framework (NasTa) presented in chapter 4. This chapter describes the evaluation procedure of NasTa which involves the creation of an optimization dataset to use it in the optimization parameters process, evaluation of the Arabic short text algorithms created in the first and second phase of the NasTa framework development process and finally, comparing the performance of the Arabic short text algorithms to determine whether a combination should be used profitably in NasTa framework.

Chapter 7 concludes the thesis, highlights its contributions and suggest some new research directions for future work.

Chapter 2

Issues of Arabic Natural Language Processing

2.1 Introduction

This chapter will review the characteristics of the Arabic language and their influence on the automatic processing of Arabic, including Arabic script, Arabic word structure, sentence structure and Parts of Speech (POS) classifications. Word structure or morphology relates to regulations, rules, and processes regarding the meaningful units of language, in terms of whether these units are words or parts of words, such as different type of affixes (Ryding, 2005). The structure of the Arabic word is considered highly systematic in that it exhibits rigorous and elegant logic. This is explained in some detail in this chapter. The Arabic POS classification dilemma and its influence on Arabic pre-processing techniques including the morphological analyser, the POS taggers and the text parser are also discussed in this chapter.

There is a review of the two well-known Arabic morphological analysers which have been developed to deal with the internal structure of Arabic words and the current Arabic POS taggers which were developed to assign the POS of each word in the text. Word Sense Disambiguation (WSD) in general is described together with the main strategies which have been utilized to perform the WSD. The main source of ambiguity in the Arabic language is explained and current algorithms developed to perform the Arabic WSD are reviewed.

2.2 Arabic Language

Arabic is a Semitic language which is spoken and written by more than 300 million people in the world. It is read by 1.4 billion Muslims as it is the Holy Quran language (Farghaly and Shaalan, 2009). Classical Arabic, the standard form of the language which is used in the Holy Quran was first spoken by Arabs over fourteen centuries ago. Its grammar and vocabulary are more complex than Modern Standard Arabic (MSA). MSA is defined as the Arabs' attempt to speak Classical Arabic (Kaye,

1972). It is the formal language of Arabic countries that is used in the education sector (e.g. public schools and universities), public speeches, media and literature. MSA contrasts with colloquial Arabic, which is less sophisticated in its grammar and vocabulary than MSA. Various dialects (colloquial Arabic) are currently spoken in different parts of the Arab world. When this language is studied, the main emphasis is always placed on classical Arabic and MSA, whilst dialects are likely to be ignored (Al-Qahtani, 2005). The version of Arabic considered in this thesis is that of MSA, the language which is universally understood by Arabic speakers.

2.2.1 Characteristics of the Arabic Language

The Arabic language is considered a highly derivational and inflexional language which is based on a root and template to produce the language's words. This section addresses the characteristics of the Arabic language.

2.2.1.1 The Arabic Script

The Arabic script alphabet comprises two types of symbols (letters and diacritics) for the writing of words. The alphabet is made up of 28 letters which contain 25 consonants and 3 long vowels and which one writes from right to left. They comprise different shapes, resulting from their location in each word: for example initial, medial, final or stand-alone (Habash, 2010). These individual shapes have their origin in the Arabic style of writing whereby letters within a word are joined together in a cursive manner, subject to the context in which the words appear. The letters individually signify certain sounds and there is a good fit between the spelling of a word and the manner of its pronunciation (Ryding, 2005).

Three long vowels in the Arabic alphabet are written into Arabic words as part of the spelling of the word. They are represented by the letters ا ***alif,*** و ***waaw*** and ي ***yaa***. In the transformation process, words which have long vowels may change or replace these letters with each other. For example, the long vowel letter ا ***alif*** in the verb قـال (said) is replaced with the long vowel و ***waaw*** during the transformation process of the verb to يقول (say), whilst the long vowel letter ا ***alif*** of the verb بـاع (sold) is replaced by the long vowel ي ***yaa*** to become يبـيع (sell). In addition, Arabic script has

short vowels, which appear as diacritics above or below the letters. Consequently, the letter acquires its desired sound and thus a word receives its desired meaning (Elkateb et al., 2006b). For example, the word مَدرَسة means school. If the diacritics are changed to مُدرِسة the meaning changes to that of a female teacher. There are three main short vowels in Arabic (Fatha /a/ ◌َ, Damma /u/ ◌ُ, Kasra /i/ ◌ِ). Sukun ◌ْ indicates there is no diacritic to add a vowel. Additional Arabic diacritics are Nunations and Shadda. Nunations only arise in the final position of nouns, adverbs and adjectives and resemble a dual version of their corresponding short vowels (two Fatha ◌ً, two Damma ◌ٌ and two Kasra ◌ٍ). Shadda ◌ّ represents a consonant doubling diacritic (Habash, 2010). For example the word دَرَسَ (darasa) means study whilst the word دَرّسَ (da**rr**asa, double consonant **r**) means teach. In contemporary texts, the short vowels have been disappearing and readers are anticipated to fill in the missing diacritics by applying their knowledge of the language. The omission of short vowels from Arabic texts results in considerable ambiguity and poses challenges to the automatic processing of Arabic (Habash, 2010).

Another symbol used in current Arabic script is that of punctuation. The Arabic writing system uses punctuation marks which are similar to those used in European languages. Attia (2008) reported that punctuation marks have been introduced into the Arabic writing system recently to some extent which has resulted in the absence of strict punctuation rules. Arabic writers write entire paragraphs without a full stop and sentences are often connected by the coordinating conjunctions و ***wa*** and ف ***fa***. With regard to this, Daimi (2001) declared that "Arabic is distinguished by its high context sensitivity with the desire to exhibit the different synthetic coherence relations".

Arabic script does not use capitalization: as a result there is no distinction between small and capital letters in Arabic. Furthermore, Arabic script does not combine letters to generate a new sound as in English (Salem, 2009).

2.2.1.2 The Arabic Word Structure

Word structure or morphology concerns the regulation, rules, and processes of the meaningful units of language, in terms of whether these units are words or parts of

words, such as different kinds of affixes (Ryding, 2005). The Arabic word structure (morphology) is considered highly systematic in that it exhibits rigorous and elegant logic. Its theories focuse on two fundamental issues: derivational morphology describes how words are formed and inflectional morphology concerns how words vary or inflect in order to mark grammatical categories (Ryding, 2005).

Arabic words are formed based on a system of roots which mesh with patterns of vowels or patterns of consonants and vowels. The root is a sequence of 3 (occasionally also 2 or 4) consonants in a particular order which are called radicals. This bears the core meaning of Arabic words (lexical meaning). The pattern is a template of one or more vowels, or in combination with derivational affixes which have slots for root radicals, and possess grammatical meaning. The Arabic language has more than 10,000 roots and 85% of derived words are formed from 3 consonant (tri-literal) roots (Al Ameed et al., 2005).

(Ryding, 2005) stated that "the Arabic root-pattern process has evolved extensively and very productively in order to cover a vast array of meanings associated with each semantic field". For example, most of the Arabic words (in different POS) which relate to writing are formed from the root of three consonants k-t-b (writing-related) as a result of switching in patterns of vowels or patterns of vowels and consonants, as shown in figure 2.1. The produced words can function as stems for grammatical affix in the inflectional stage.

Root	كتب k-t-b R_1-R_2-R_3					
Pattern	$R_1aR_2aR_3a$	$R_1iR_2aaR_3a$	$R_1aaR_2iR_3$	$R_1iR_2aaR_3$	$maR_1R_2aR_3a$	…
Stem	kataba write (v.)	kitaaba writing (n.)	kaatib writer (n.)	kitaab book (n.)	maktaba Library (n.)	…

Figure 2.1 The formation of some Arabic words (writing related) from the root كتب k-t-b.

Inflectional morphology does not change the core meaning and part of speech of the stem but grammatical affixes are added in order to mark grammatical inflections, such as tenses (past/present), gender (masculine, feminine) and/or numbers (singular, dual (representing two entities), plural). For example, adding the suffix "ان" ***an*** to the stem "*kaatib*" (writer) produces the word "*kaatiban*" (two writers) which signifies the dual masculine.

A multiple affix can appear in a word, when particular coordinating conjunctions, prepositions and particles, the definite article, and a class of pronouns attach themselves to the words. Thus a single Arabic word can represent a complete sentence in other languages. An example of this is the Arabic word واخبرتهم which means "and I told them". This feature makes pre-processing tasks of Arabic texts very challenging as it hinders the matching of the word in Arabic text to the correct sense (correct lemma). It also poses two interesting challenges to the STSS computation: representation of the word in a short text especially for measure that calculates the similarity based on bag of words and also extraction of the semantic information from Arabic resources (described in chapter 3) directly where the Arabic words have been saved in these resources as lemmata.

2.2.1.3 Sentence Structure and Word Order

Arabic sentences have been classified as equational (verbless) sentences and verbal sentences (Ryding, 2005, Attia, 2008). The equational sentence is a sentence without a verb and its structure consists of the subject and predicate. The subject is a noun phrase whilst the predicate can be a noun phrase, adjective phrase, adverb phrase or prepositional phrase. An example of this as follows:

1. اخي مهندس / My brother is an engineer. In this example the first word اخي (my brother is a subject (noun) and the second مهندس (engineer) is a noun phrase predicate.
2. الطالب ذكي / The student is intelligent. The word الطالب (the student) in this sentence is a subject and ذكي (intelligent) is an adjective phrase predicate.

A verbal sentence is a regular sentence in Arabic and its structure consists of the verb, subject and object. The verbal sentence is considered syntactically flexible and has a relatively free word order (Attia, 2008). Every different order, the Subject-Verb -Object (SVO), VSO, and VOS are acceptable sentence structures in MSA. The English sentence (the man bought a car) can be written in Arabic in three ways as follows:

1. VSO order, اشترى الرجل سيارة / bought the man a car
2. SVO order, الرجل اشترى سيارة / the man bought a car
3. VOS order, اشترى سيارة الرجل / bought a car the man

This feature poses a challenge for many Arabic applications such as machine translation (Salem, 2009), Arabic parsing (Attia, 2008) which increases the ambiguity and conversational agent (Hijjawi et al., 2014) which increases the complexity in terms of the actual understanding of Arabic sentences. For the work in this thesis, the Arabic STSS measure cannot take advantage of word order which contributes to English STSS measures.

In addition, MSA is a pro-drop language whereby the subject pronoun of a verb in a sentence is dropped and recovered later by convention. For example, the Arabic sentence اشترى سيارة (bought the car) is equivalent to هو اشترى سيارة (he bought the car) (Diab et al., 2007 and Farghaly and Shaalan, 2009).

2.2.1.4 Parts of Speech

Compared with English and other European languages, Arabic enjoys a longer millennium-wide tradition of scholarly research relating to its grammatical description. The order established by the Arabic grammarian Sibawaihi, approximately fourteen hundred years ago is the method most frequently followed in traditional grammatical studies. In his renowned book Al-Kitab, Sibawaihi (1966) commences by classifying the Arabic POS into nouns, verbs and particles. The verb indicates an action and tenses that apply; Nouns which include people names, places, or objects have no tenses; the particle requires that it is joined by a verb or a noun or

both in order to be understood (Sawalha, 2011). This classification is still used today and is regarded as the Arabic grammar's leading principle (Suleiman, 1990).

The classification of POS is not listed in Arabic dictionaries whilst the structure of Arabic grammar books is subject to the division of POS into nouns, verbs, and particles. Wright (1896/2005), for example, applied the term noun as an umbrella etymology covering six types which include nouns, adjectives, numeral adjectives, demonstrative pronouns, relative pronouns and personal pronouns. He also divided particles into: prepositions, adverbs, conjunctions and interjections (Attia, 2008).

In the literature of modern Arabic linguistics, Suleiman (1990) carefully analysed the work of the earliest Arabic grammar theoretician (Sibawaih), in his book "Al-Kitab" and refuted his tripartite classification of Arabic POS. The main thrust of his argument was that no empirical or reasonable evidence was given by Sibawaih to support his theory that the Arabic POS are exclusively classified into nouns, verbs and particles. This was also the opinion of Attia (2008) who acknowledged that classifying the Arabic POS in the traditional manner into nouns, verbs and particles is insufficient for providing a complete computational grammar. This is supported by Sawalha (2011) who observed that the tripartite classification of Arabic POS by Sibawaihi does not pay sufficient attention to word structure (morphology).

This issue affects the Arabic text pre-processing techniques including morphological analysis (analysis of the Arabic words), POS tagging (assigning a grammatical category label to each word in a text) and text parsing (assigning a syntactic structure to a group of words). The morphological analyser is considered a precondition for the POS tagger and the text parser which provides them with the most important information they need. A considerable number of morphological analysers (used to analyse the Arabic words) continue to be influenced by the tripartite Arabic POS classification (Attia, 2008). A good example of this restricted point of view is the Xerox Arabic morphological analyser (Beesley, 2001). In these morphologies, Arabic words are strictly classified into verbs, nouns (including adjectives and adverbs) and particles allowing for no additional categorical description to be used and thus making them unsuitable to serve a POS tagger and a syntactic parser (Attia, 2008).

2.3 Arabic Morphological Analysers

Two principal strategies apply for developing Arabic morphologies. They are dependent on the level of analysis as follows:

1. Root-based morphologies: the analysis of Arabic words based on the system of roots and patterns as well as concatenations.

2. Stem-based morphologies: the analysis of the Arabic words at the stem level with the use of regular concatenation. The stem is considered the least marked form of the uninflected word and has no suffixes or prefixes. It is normally the perfective, third person, singular verb in MSA whilst nouns and adjectives appear in the form of singular indefinite.

Many morphological analysers were developed for Arabic but only some of them are available for purposes of research and evaluation, the remainder are proprietary commercial applications (Attia, 2008). The known analysers include Buckwalter Arabic Morphological Analyser (BAMA) (Buckwalter, 2002), Xerox Arabic Morphological Analyser (Beesley, 2001), Sakhr (Chalabi, 2004), Diinar (Dichy and Hassoun, 1998), and Morfix (Kamir et al., 2002). The best known are the first quoted analysers which are well documented and are available for researchers to evaluate (Attia, 2008). Each will now be reviewed.

2.3.1 Buckwalter Arabic Morphological Analyser (BAMA)

BAMA Morphology is regularly found in the literature and is believed to be the "most respected lexical resource of its kind" (Hajic et al., 2005). BAMA contains 38,600 Arabic lemmata and has been developed as a main database of Arabic word forms which interact with two concatenation databases. Arabic words are viewed as a concatenation of three regions: a prefix, a stem and a suffix. The prefix and suffix regions can be null. Prefix and suffix lexicon entries cover all possible concatenations of Arabic prefixes and suffixes, respectively. Each word's form is inputted separately. The stem is taken as the base, and information about the root is

also given. BAMA acts to verify the probable existence of each part in the three dictionaries and is deemed acceptable if the prefix and suffix are null.

There are three compatibility tables in BAMA which are accessed after the word is divided into its prefix, suffix and stem and a match for each is located in the lexicons. Verification of a compatible combination is subsequently undertaken by means of the compatibility tables. Successful verification indicates correct spelling of the word. The vowel marks are reconstructed by BAMA. An English glossary is provided and every possible combination of stems and affixes for a word is made available. All stems that have a similar meaning are grouped together by BAMA and then linked to a lemma ID. A Modern Written Arabic Dictionary (Wehr, 1979) was taken by Buckwalter as his reference source.

Arabic words are classified by BAMA based on the modern POS classification. There still remain, however, traces of generalizations in the large number of adjectives categorised to be nouns and particles are deemed to be function words (Attia, 2008).

2.3.2 Xerox Arabic Morphological Analysis and Generation

According to (Dichy and Fargaly, 2003), Xerox Morphology is held to be a system based on "solid and innovative finite-state technology". It is a mathematical model which was used for the design of programs that can be signified via states and the transition between them (Attia, 2008). The machine has been adapted to the Xerox finite-state format. Beesley (2001) presented a description of this system which is believed to be more appropriate for the carrying out of morphological analysis. The approach of root-and-pattern is adopted by this morphology. 4,930 roots and 400 patterns are included, with 90,000 stems effectively generated. The advantage of using it being the fact that it is rule based and has a large coverage. Vowel marks are also reconstructed and an English glossary provided for each word.

It is subject to POS classification specifications, thus making it unsuitable to serve a syntactic parser as words are classified only into Verbs, Nouns (including adjectives

and adverbs) and Particles (Attia, 2008). A principle disadvantage of Xerox morphology is the increased rate of ambiguity. Attia (2008) stated that, on account of the fact that the system gives so many analyses for most words, including many spurious ones due to the previous mentioned factor, it suffers from a very high level of ambiguity.

It was decided to use the BAMA Arabic morphological analyser in this study to obtain the lemma of each Arabic word in the short text as BAMA is freely downloadable as a java package whereas the Xerox system is a web based analyser. BAMA classifies words utilising modern POS classification and takes the stem as its base form. By contrast, Xerox is based on traditional POS and utilises root–pattern which increases the ambiguity, resulting in an increase in the number of solutions, which Xerox morphology provides for most words.

2.4 Arabic Part of Speech Taggers

POS tagging is the process of assigning a word class (grammatical category label) to each word in a text and is automatically performed using the POS tagger technique. The set of all grammatical category labels used in the tagging process is known as a POS tag set. The development of Arabic POS tagging has started recently and various techniques have been employed to resolve the problem of Arabic POS tagging.

2.4.1 Stanford Part-Of-Speech tagger

Stanford University originally developed this tagger (Stanford tagger) to apply to the English language (Toutanova and Manning, 2000). A further, improved version was presented which adds support for different languages together with improved speed and usage for English which was described by (Toutanovaet al., 2003).

The tagger is built based on the model of maximum-entropy. The maximum entropy intuition is to create a distribution through the continuous addition of features (Jurafsky and Martin, 2000). The term 'features' refers in this context to the

constraints which come into being when the tagger is trained, e.g. syntactical and morphological features. The total distribution contains the constraints that are added by each feature. (Jurafsky and Martin, 2000) provide further information. The most recent version includes trained models for the Chinese, German and Arabic languages. According to the authors in the README file, the tagger has 96.50% accuracy in Arabic. The tagger concentrates on the training part of the Penn Arabic Treebank (PATB) with a smaller POS tag set which makes it harder to allocate a "wrong" tag, thus contributing to the high level of accuracy. Examples from the set of POS tags used include (NN- Noun single, NNS - Noun plural, DT- Determiner, JJ- Adjective, VBD – Verb past tense, ect.)

2.4.2 Khoja Arabic Part-Of-Speech Tagger

In Khoja (2001) a combined statistical and rule-based method were proven to yield the best results from the various combinations experimented with. A set of 177 POS tags is used by the tagger which originates from Arabic traditional grammatical theory. This set consists of 103 tags for nouns, 57 for verbs, 9 particles, 7 residual and 1 for punctuation. The rule based method involves the development of a knowledge base of rules which has been written by linguists as a means of defining accurately how and where to allocate the various POS tags. The statistical based method involves the building of a trainable model and the usage of a tagged corpus for estimating its parameters. Once accomplished, the tagger can be employed to automatically tag other texts.

The Khoja testing phase used four different corpora. The largest corpus, amounting to approximately 59,000 words, was employed to train the tagger and create a number of lexicons, which were used to tag the test set. One of these lexicons listed each word jointly with all possible tags which were obtained in the corpus. In the initial stages of the tagging, each word was looked up in the lexicon and all possible tags for the word were identified in the lexicon. A stemming process was performed for any word that was not found in the lexicon. In all, the accuracy achieved by Khoja tagger was around 90%.

2.4.3 Automatic Tagging of Arabic Text

Another Arabic POS tagger was introduced by (Diab et al., 2004) using a learning algorithm known as the Support Vector Machine (SVM). This is a supervised machine learning algorithm which is robust and can handle a big number of features. It enjoys a good general performance. A number of features were drawn from a predefined linguistic context with the tagger designed to predict the class of a token. Arabic TreeBank was used to train the tagger and the data in the Arabic TreeBank was transliterated into Latin based ASCII characters by means of the Buckwalter transliteration scheme. A set of 24 POS tags was used by this tagger to achieve a level of accuracy of 95%. This set of 24 tags known as collapsed tags set was manually selected from the set of 135 tags created by Buckwalter (2002) to use with Arabic morphological analyser.

2.4.4 Hybrid Method for Tagging Arabic Text

A hybrid method was presented by Tlili-Guiassa (2006) for tagging Arabic text by combining a rule-based method and a memory-based machine learning method. In the simple memory-based learning method, appropriate examples are given for memory retention and the similarity between memory examples and new examples resulting in the prediction of new examples. The tagger determines the word x POS tag by searching for the k nearest neighbours and selecting the neighbour with the highest frequency of occurrence. In the testing phase, the tagger used a corpus containing texts which have been drawn from first stage educational books and Qur'anic text that has been tagged through the use of a small tag set. The POS tag set used by this tagger is the set of POS tags derived from Khoja's tagger resulting in a performance of 85%.

(Sawalha, 2011) drew attention to the fact that most of the Arabic POS taggers were developed by NLP research groups for their own internal use only. The reported taggers used different sets of POS tags and evaluated using different test corpora. Of all the Arabic taggers, Stanford enjoys the highest performance score. Moreover, it is the only tagger that is freely available for download by researchers and therefore

subject to independent validation. Consequently, it has been adopted in this study for tagging each word in the Arabic short texts.

2.5 Arabic Parsers

Parsing is the process of assigning a syntactic structure to a group of words and is automatically done using the text parser technique. This technique has been used in variety of NLP applications such as automatic summarization and machine translation (Habash, 2010). Several parsers have been used for parsing Arabic text such as the Stanford parser (Klein and Manning, 2003), the Bikel parser (Bikel, 2002), Malt parser (Nivre et al., 2007), an Arabic Slot Grammar parser (McCord and Cavalli-Sforza, 2007) and a Rule based parser (Attia, 2008). The parser presented by (Attia, 2008) on the bases of the f-structure discussed later will be used in this research in order to manage the syntactical flexibility feature for the MSA. Therefore this parser is presented with more details in this section.

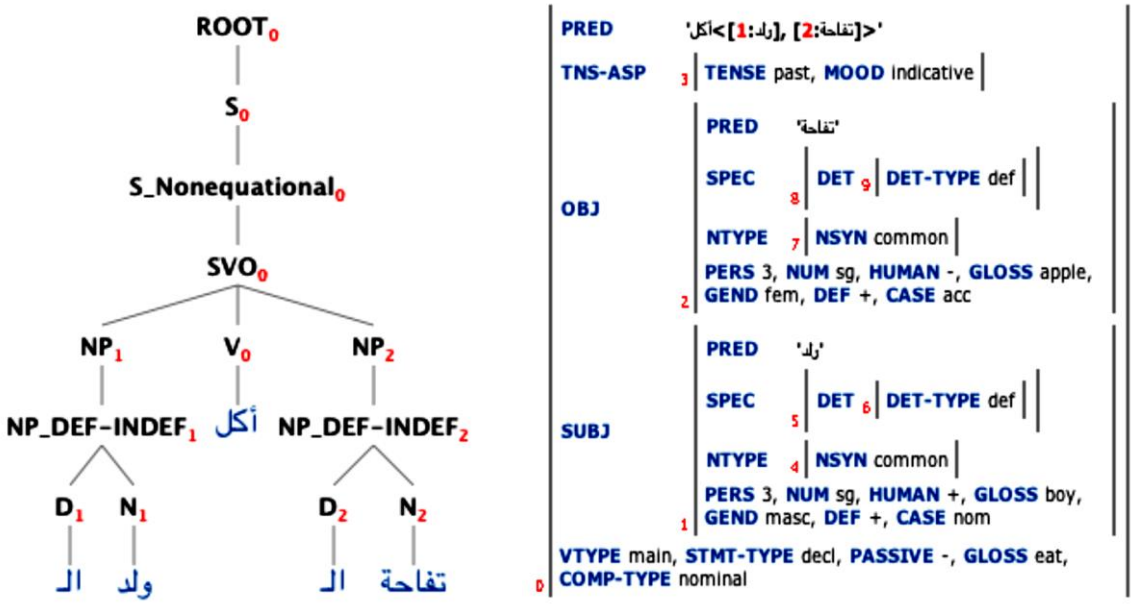
An Arabic parser was developed by Attia (2008), who created it within a framework called a Lexical Functional Grammar (LFG) (Bresnan, 2001). This was undertaken by means of the Parallel Grammar (ParGram) Project's (Butt et al., 2002) formalisms, tools and common inventory.

The aim of ParGram project is to provide full syntactic representation for many languages within the LFG framework. The project utilises the Xerox Linguistic Environment (XLE) as a platform that was built by Palo Alto Research Centre (PARC) in order to write grammar rules and lexical entries using LFG formalisms. The platform consists of three components suitable for creating a machine translation system which include a parser, transfer and generator (Attia, 2008).

After being supplied with enough rules and lexical entries, the system analyses (parses) sentences and gives both the functional-structure (f-structure) and constituent-structure (c-structure) representation for each sentence. The c-structure is defined as a phrase structure tree which encodes consistency (dominance) and precedence (surface order) for each sentence (Attia, 2008). The f-structure represents

a level of abstraction which is high enough for capturing parallelism amongst different languages and reduces cross linguistic syntactic differences. It gives information on grammatical functions of words such as (subject and object) and morpho-syntactic features such as tense, gender, number and person (Attia, 2008).

The grammar rules and notations for MSA were written using the XLE platform.
The results of the pre-processing stages in the XLE system include grammatical category, essential morphological information and the morpho-syntactic features for each word. A set of Arabic rules, notations and constraints are employed to analyse the Arabic sentences by the XLE parser. The main results obtained by the XLE parser after parsing the Arabic sentences are the f-structure and the c-structure for the input sentence. This parser is available online at http://iness.uib.no/iness/xle-web which allows input of an Arabic sentence and gives the f-structure as output. Figure 2.2 shows the c-structure and f-structure for the Arabic sentence الولد اكل التفاحة "the boy ate the apple" which is selected from (Attia, 2008).



(a) **(b)**

Figure 2.2 the Rule based parser output a: c-structure, b: f-structure.

The output of this parser (f-structure) will be used in this thesis to manage the syntactical flexibility feature for MSA; the consequence of this is described in chapter 4.

2.6 Word Sense Disambiguation

Each individual word can possess several possible meanings, a process called Polysemy. The human being is able to ascertain the intentional meaning of a word used by another person in conversation and in writing. These possible meanings are known as senses and computers find it more difficult than human beings to comprehend the intentional meaning of a word in a given context. As a result, several algorithms for Word Sense Disambiguation (WSD) have been developed to perform this task. This acts to identify the correct sense of a particular word based on the context in which it appears (Navigli, 2009).WSD has been a problem in Computational Linguistics for a long time and impacts significantly on many real-world applications, such as machine translation, information extraction, and information retrieval.

WSD was originally considered as a part of Machine Translation in the late 1940's, when the use of computer software to undertake translations of one language into another was under consideration. However, it was rapidly evident that it presented a serious challenge and, indeed, WSD was subject to various attempts in the 1970's to resolve the problem by means of the use of Artificial Intelligence techniques. A turning point in WSD research was achieved in the 1980's when the large scale lexical resources released allowed for automatic extraction of knowledge (e.g. Wilks et al., 1990). Statistical and machine learning techniques were heavily used to perform WSD in the late 1990's.

The generic WSD task can be distinguished by means of two distinct tasks, which are the target word and all word. In target word (or lexical sample), a single ambiguous word is subject to being disambiguated in a given context. All-words WSD is a more general method which includes the disambiguation of all content words (nouns, adjectives, verbs and adverbs) in a text simultaneously.

The methods proposed to tackle WSD generally employ the context in which the ambiguous word occurs in order to disambiguate it, and use external knowledge resources to extract the context information. The fundamental component of WSD is the knowledge resource which can be partitioned into two types of resources structured and unstructured. Structured resources comprise thesauri, machine readable dictionary and ontologies such as Roget's International Thesaurus (Roget, 1911), Collins Cobuild English Dictionary (Sinclair, 2001) and English wordnet (Miller, 1995), respectively whilst unstructured resources comprise raw corpora and sense-annotated corpora. An example of raw corpora is the Brown Corpus (Kucera and Francis, 1967), which is published in 1961 in the United States and consists of one million word collection of texts. A further example of unstructured texts is the largest sense-tagged corpus known as SemCor (Miller et al., 1993), which contains 352 texts tagged with approximately 234,000 sense annotations. The reported knowledge resources are described with more details in (Ide and Vèronis, 1998).

Several approaches have been proposed to perform WSD which can be categorised into 4 groups.

The supervised approach is popular due to its higher performance which trains a supervised learning algorithm using a large amount of manually annotated training data. Several machine learning algorithms have been used in supervised WSD such as decision trees, neural networks, Naive Bayes classifier, decision lists, support vector and instance base learning. A detailed description of each of these algorithms was given in Navigli (2009). The supervised learning algorithm trains a classifier using a set of labelled training data and generating a statistical model. This model is applied to a set of unlabelled test data to decide the appropriate sense for each ambiguous word. One of the significant disadvantages of this approach is that it requires a large amount of manually annotated training data which is usually created by humans. Unfortunately, human sense-tagging is labour intensive and time consuming (Navigli, 2009). It requires a human expert to be very familiar with each word's definition. In the lexical sample task, a human manually tags each occurrence of a single word (target word) in a text whilst in the case of all-words task, a human manually annotates all content words (nouns, verbs, adjectives and adverbs) in a text.

The limitation of the supervised approach leads to the use of semi-supervised and unsupervised approaches. The semi-supervised approach requires very small set of labelled training data (called as seed data) as in bootstrapping processes which trains the classifier with a small tagged corpus (Yarowsky, 1995) and then applies the classifier to annotate a set of untagged examples selected randomly from a large set of unlabelled data. The results of this step (a new set of annotated examples) are added to the small tagged corpus. This process is repeated with new sets of untagged examples from the large set of unlabelled data until reaching a specific threshold. Some semi-supervised approaches used the word-aligned bilingual corpus as seed data (Ng et al., 2003).

The unsupervised approach does not rely on a labelled training data and includes clustering which performs the WSD based on the notion that "the same sense of a word will have similar neighbouring words" (Navigli, 2009). Therefore, clusters of words are created based on the adjacent words (Lin, 1998a). All the described approaches were reviewed in Navigli (2009) and Ide and Vèronis (1998) and acknowledged that the supervised approaches with sufficient annotated training data outperformed the unsupervised approaches. However the unavailability of such sufficient data leads to the use of unsupervised approaches for wide coverage WSD.

The knowledge based approach typically utilizes external knowledge to perform WSD and does not require any manually labelled training data. It is considered the most promising approach for WSD due to the availability of the external knowledge such as the dictionaries, thesauri, lexical databases and ontologies (such as wordnet, which is increasingly enriched) (Pedersen et al., 2005). Several methods have been proposed to perform WSD by exploiting the knowledge resources structure. A simple knowledge based approach is the gloss overlap or known as Lesk algorithm (Lesk, 1986) which performs WSD by calculating the word overlap between the target word senses' definitions (dictionary definition) and the definitions of the senses of the adjacent words in the sentence. The sense of the target word that has a highest overlap is assigned as the appropriate sense. (Banerjee and Pedersen, 2003) extended the gloss overlap using English WordNet instead of the dictionary which exploited the different relationships that connect the concepts in WordNet. The structural

approach is another knowledge based approach which performs WSD using a semantic relatedness or similarity measure (Pedersen et al., 2005).

2.6.1 Arabic Word Sense Disambiguation

Arabic has a higher degree of ambiguity due to a complexity in the Arabic writing system. The reason is that the absence of short vowel representation in MSA causes an increase in homographs (words have same spelling but different pronunciations, usually with different senses) (Habash, 2010). For example, the Arabic word بر could mean three different nouns, بَرّ (land) or بُرّ (wheat) and بِرّ (reverence). Also like other natural languages, most Arabic words are polysemous (word with one pronunciation has multiple senses). For example, the Arabic word جبن which mean cheese or cowardice. Both homograph and polysemy are cases of WSD. Maamouri and Bies (2010) illustrate that the average number of ambiguities for a word in most languages was 2.3, while in MSA it was 19.2. This section will review the current algorithms developed to perform Arabic WSD.

2.6.1.1 An Unsupervised Approach for Bootstrapping Arabic Sense Tagging

An unsupervised machine learning approach was presented by (Diab, 2004) for Arabic word sense tagging, known as "bootstrap". This approach used a parallel Arabic-English corpus for the annotation of the Arabic text (focusing on nouns) which utilized the cross-linguistic correspondence to characterize word meanings. The words in the Arabic text were annotated based on the notion that words in the first language were translated into the same word in the second language then the first language words are semantically similar. The Arabic words were annotated with their meaning definition using the English WordNet.

A word aligned parallel corpus was taken as input by the proposed algorithm (for each Arabic word an English word was collected with). All English words that were translated into the same Arabic word (same orthographic form) were collected from the corpus and grouped into clusters. For each word in the cluster, all possible senses were determined using English wordnet and the appropriate sense was assigned

following the same algorithm used by (Resnik, 1999) to disambiguate the group of English nouns. In the final step, the proposed algorithm projected the chosen sense tags for English words to the corresponding translation words in Arabic. In the test experiment, an all word test corpus (SENSEVAL2) was used for English whilst machine translation systems were used to generate the Arabic due to lack of an Arabic test corpus. The proposed algorithm achieved 64.5% in precision and 53% in recall on the SENSEVAL2 English All Word task whilst 90% of the Arabic evaluated data items were accurately tagged by the proposed algorithm based on Arabic native judgment (annotations and ratings).

2.6.1.2 Naïve Bayes Classifier for AWSD

A supervised approach was applied by (Elmougy et al., 2008) for Arabic language to disambiguate a single word in a text which used a Naïve Bayes classifier with the rooting algorithm to solve the ambiguity of Arabic words. A Naïve Bayes classifier relied on the computation of the conditional probability of occurrence of each sense of the ambiguous word in the given context. The sense of the ambiguous word which maximizes its conditional probability is chosen by Naïve Bayes classifier as a correct sense in context. The Naïve Bayesian classifier can be represented by the following formula:

$$P(s_i | f_1, f_2, \dots, f_n) = p(s_i) \prod_{j=0}^{m} p(f_j | s_i) \qquad 2.1$$

Where s_i represents the ambiguous word sense, f_j represents the features that used for describing the context in which the ambiguous word appear and m represents the number of features. The probability of sense p (s_i) and the conditional probabilities p (f_j | s_i) are estimated based on the relative occurrence frequencies of feature f_j and sense s_i in the training set.

Elmougy's algorithm started the disambiguation with two pre-processing steps which were applied to eliminate the stop words and to extract the root of each Arabic word. The AlShalabi stemmer (Al Shalabi et al., 2003) was used for the root extraction which analyse the Arabic words based on the system of roots and patterns. In the training phase, the training set was collected using the net and dictionary whereby

ten training samples were collected for each predefined ambiguous word. The output of the training step was used by the disambiguation algorithm to calculate the score of each ambiguous word sense and to assign the correct sense for a given word in the set of testing samples. The testing set was also collected from the World Wide Web. This algorithm achieved a rate of precision of 73% and the authors claimed that using the root extraction algorithm with Naïve Bayes classifier improved the accuracy and also reduced the dimensionality of the training samples.

2.6.1.3 Corpora based Approach for Arabic/English Word Translation Disambiguation

An Arabic/ English word translation disambiguation algorithm was proposed by (Ahmed and Nurnberger, 2009) based on Naive Bayes classifier. The proposed algorithm disambiguated the user translated query to assign a most appropriate word translation based on statistical co-occurrence with utilizing a large bilingual corpus.

The proposed algorithm used an Arabic/ English parallel corpus for training and testing phases. This corpus contains 8,439 Arabic stories with their English translations totalling 2 Million Arabic words with 2.5 Million English words.

The lemma of each word in the user query was extracted using BAMA Arabic morphological analyser and then each word in the user query was translated to English. All possible English translations were determined for each word in the user query and stored in the sense inventory array. The Naïve Bayes classifier then started the disambiguated process of the ambiguous query word (as described in section 2.5.1.2) and the sense matching the highest number of features was assigned as a most appropriate word translation. The evaluation process used Arabic sentences from the bilingual corpus as a user query. This algorithm using inflectional form (lemma) achieved 93% in precision compared with the same manually selected senses in both cases whilst 68% achieved using the basic word form.

2.6.1.4 Lexical Disambiguation of Arabic Language: An Experimental Study

An experimental study was presented by (Merhben et al. 2012) using three supervised algorithms to perform Arabic word sense disambiguation. These are the Naïve Bayes algorithm, the K Nearest Neighbour and the decision list (Navigli, 2009) which are considered the most popular and the highest performing supervised algorithms in WSD.

The experiment started the disambiguation with two pre-processing steps which were applied to eliminate the stop words and to extract the root of each Arabic word. Khoja stemmer (Khoja et al., 1999) was used for the root extraction which analyse the Arabic words based on the system of roots and patterns. In the training phase, a non-annotated corpus produced by (Al-Sulaiti and Atwell, 2006) was used and four Arabic annotators tagged the 50 ambiguous words (from the corpus) by their senses.

For the 50 ambiguous words selected, the K Nearest Neighbour algorithm achieved the highest performance of 52.02 % among others and the stemming increased the precision for the three algorithms between 9% and 21%.

2.6.1.5 A Semi-Supervised Method for AWSD Using a Weighted Directed Graph

A semi-supervised method was proposed by (Merhbene et al. 2013) which combined a supervised method and an unsupervised method for disambiguating a single Arabic word in a text. The proposed algorithm consisted of three steps.

Step 1 presented a method that was used to cluster the Arabic sentences containing the ambiguous word. This step used the Arabic WordNet (AWN) (Elkateb et al., 2006a) to extract the glosses (definition) and synonyms of the ambiguous words. Also the corpus (collected by authors from newspaper articles, which counts 123,854,642 words) was used to collect sentences containing the ambiguous words. For each sense of the ambiguous word, a sense cluster was produced by grouping the sentences that represented the meaning of this sense. These clusters were then used to construct the semantic trees. Accordingly, the sentences in each cluster were

transformed to binary trees which consisted of nodes, edges, root (represents the ambiguous word), right children and left children. All the obtained trees were merged with the corresponding sentences listed in the same cluster.

Step 2 included the construction of a weighted directed graph and called a matching step. The weighted directed graph was constructed by matching the original sentence tree with the produced semantic trees of each sense. Edges weighted were added between the nodes of the original sentence tree and the semantic tree nodes using three collocation measures. These measures are the T-test, the Mutual Information and the Chi-Square (Maning and Schütze, 1999). The weighted directed graph was employed to determine the closest semantic tree to the sentence tree being disambiguated, using a score measure which created based on the collocation measures.

Step 3 presented a voting procedure which was used to assign the correct sense to the ambiguous word. This procedure ranked the collocation measure in accordance with the correct attribution of the given sense. The sense obtaining a highest rank from the collocation measure was assigned to the ambiguous word.

The test process used a manually tagged (by Arabic annotators) test data of 4,582 samples containing 127 Arabic ambiguous words. The algorithm achieved (83%) in a recall and precision.

2.7 Category Norms

A category norm is defined as a set of words within the same theme, listed by frequency, which is created as responses by human participants to a specific category (Battig and Montague, 1969). The words in each category are more similar to each other than to the words of other categories. Battig and Montague (1969) created the original category norms and their work is considered the best-established set which is used in many projects, for example (Marsh et al., 2008, Caramazza and Shelton, 1998). The success of these categories may be attributed to the authors' objective that "these category norms may differ from numerous other similar normative projects because of our primary concern with making them as useful as possible for other researchers". A follow up study was carried out by (Van Overschelde et al., 2004) and reported that the category norms of Battig and Montague have been

employed in over 1600 projects which were published in over 200 different journals. Examples of these categories are:

1 A precious stone: Diamond, ruby, emerald, sapphire, pearl, gem
5 A metal: steel, iron, silver, copper, gold, aluminium, platinum

The membership data of 56 Battig and Montague categories were updated by (Van Overschelde et al., 2004) and also 14 new categories were added resulting in 70 category norms. Battig (Battig, 1979) placed emphasis on the verbal material's importance for the research community, together with perceived difficulties in obtaining the necessary funding to produce them. For example, (Van Overschelde et al., 2004) created 70 category norms using a sample of 600 participants per category and the participant's responses for each category were typed into the computer rather than handwritten.

There is a need for constructing materials in Chapter 5 and 6 of this thesis for the creation of data sets to enable the evaluation of both the Arabic noun similarity measure and the Arabic short text similarity measure. This process requires employing categories like Battig and Montague. However, they cannot simply be adopted because the content of the category norms differ from one language to other on the basis of the culture (Yoon et al., 2004).

2.8 Conclusions

This chapter has described the characteristics of the Arabic language, including Arabic script, morphology, sentence structure and POS classifications. It has been shown that the characteristics of such a rich language pose significant challenges to automatic processing which included missing diacritics, complex internal word structure, relatively free word order, pro-drop language and different POS classifications.

Two well-known Arabic morphological analysers have been reviewed. The BAMA morphological analyser was deemed the most suitable for adoption in this study.

Current Arabic POS taggers have been reviewed based on the algorithm utilized, the training and testing resources used, tag set and the accuracy achieved. It was decided that Stanford POS tagger for Arabic is the most appropriate for use in this research due to its accuracy and availability.

Explanations of WSD in general have been presented together with the main strategies used to perform the WSD. It has shown that the supervised approaches are popular due to its higher performance but the knowledge base approaches are the most promising due to the availability of the external knowledge. Details of current Arabic WSD algorithms have been reviewed as regards the methodology used, knowledge resource exploited and the accuracy achieved by each algorithm. This review demonstrates that the majority of Arabic WSD algorithms were developed for single word WSD task only and no implementation is available on the web for adoption of them by researchers such as a freely available package of WSD for English. An Arabic WSD algorithm to disambiguate all words in the Arabic short texts will be presented in chapter 4 of this thesis which is based on the knowledge base approach.

Chapter 3

Semantic Similarity

3.1 Introduction

Semantic similarity is an essential component of numerous applications in fields such as natural language processing, linguistics and psychology. Semantic Similarity is believed to be a widely understood concept. In word semantic similarity study, Miller and Charles (1991) wrote: ". . . subjects accept instructions to judge similarity of meaning as if they understood immediately what is being requested, then make their judgments rapidly with no apparent difficulty." This view has been supported by other researchers such as Resnik (1999) who noted that similarity is generally treated as a property which is characterised by human perception and intuition.

Different semantic types were discussed by Frawley (1992) with respect to two mechanisms. These are the detection of similarities and differences. Jackendoff (1983) claims that the synonym, redundancy and paraphrase semantic relations derive from judgements of likeness while the semantic relations of antonymy, inconsistency and contradiction arise from judgements of difference.

Fellbaum (1998) stated that words and texts are considered semantically related when a relationship exists between their meaning. A pair of terms can be semantically related by means of lexical relationships such as hyponymy (father, parent), synonymy (gem, jewel), and antonymy (local, international), and also by functional relationships such as (pen, paper), associative relations (winter, cold), temporal relation (World War II, 1945) for instance. Semantic relations which apply at other, higher levels, such as in phrases, sentences and documents, are subject to analysis based on their meaning within the texts.

Similarity-based research can play a crucial role in the development of the performance of the bulk of applications relying on it (Feng et al., 2008). Examples comprise word sense disambiguation (Sinha and Mihalcea, 2007), information retrieval (Hliaoutakis et al., 2006), semantic search (to find pictures, documents, jobs

and videos) (Aytar et al., 2008), information extraction (Poon and Domingos, 2007), question answering (De Boni and Manandhar, 2003), machine and conversational agents (O'Shea K. et al., 2010).

Semantic similarity studies have generally focused on one of three levels of detail: individual words, short texts or complete documents. In relation to the work in this thesis, this chapter focuses only on the word and short text semantic similarity. To the best of our knowledge, there is no prior work that has been reported on Arabic word semantic similarity measures or on Arabic short text semantic similarity measures. However, related work on English word and short text similarity measures provides a starting point. This chapter reviews the current state of the English word and short text semantic similarity measures based on the information sources they exploit. The current states of the English datasets that are used to identify the quality of the computational semantic similarity (word and short text) with the challenges of the creation of these datasets are also included. Finally the Arabic knowledge resources that support semantic similarity are reviewed based on their availability.

3.2 Word Semantic Similarity Measures

Assessing semantic similarity between two words is frequently represented by similarity between concepts associated with the compared words. Interest in automatic word semantic similarity started in 1960s, particularly for the English language. Since then, a number of algorithms have been proposed using a variety of approaches which can generally be viewed in terms of the information source they exploit: path based approaches and information theory based approaches (Meng et al., 2014).

Path based approaches can also be called Edge counting-based or Dictionary / Thesaurus based approaches (Li et al., 2003) which typically use the semantic information derived from hierarchical knowledge bases to compute the word semantic similarity. Rada et al. (1989) utilized the minimum path length connecting the concepts containing the compared words as a measure for calculating the similarity of words. This was undertaken by finding the meeting point known as the

Lowest Common Subsumer (LCS) which is the most specific concept in the hierarchy that subsumes the two concepts, followed by calculating the path distance between them through it. This proposed measure calculated the similarity of medical terms using a medical taxonomy known as MeSH. Their work is considered as the basis of edge counting-based methods. A similar kind of method was proposed by Leacock and Chodorow (1998) for measuring the word similarity using the English WordNet (Miller, 1995) taxonomy. The similarity of compared words was calculated based on the shortest path length between the compared words taking into consideration the maximum depth of the taxonomy.

Wu and Palmer (1994) proposed an algorithm to calculate the word similarity using the depth of the LCS and the path lengths (the number of nodes) between the compared concepts and the LCS. The proposed algorithm was used in a machine translation system (translating English verbs to Chinese).

Li et al. (2003) presented different strategies to calculate the semantic similarity using multiple information sources, which are the shortest path length, depth and local density. Li indicated that the reported similarity measures either used the information source directly as a metric of word similarity or utilized a particular information source without consideration being given to the contribution of others. The authors claimed that the information sources should be properly processed and combined in order to attain a good measure of word semantic similarity. The strategy that obtained the best result combined the shortest path and depth nonlinearly. The result of this measure significantly outperformed previously reported word similarity measures. In this measure, the similarity increased with respect to depth of the LCS (proportional to depth of the LCS) and decreased (inversely proportional) with the path length between concepts.

The information theory or corpus based approaches principally use the frequency of a word's occurrence to calculate the word semantic similarity using statistical information derived from a large corpus. Resnik's measure (1995) is the first to combine ontology and a corpus together. The proposed measure defined the semantic similarity of the compared concepts as the information content of the LCS that subsumed the compared concepts in the taxonomy hierarchy. The Information

Content (IC) of a LCS relies on the probability (*p*) of encountering an instance of the LCS in a corpus which is calculated using the following formula.

$$IC = - \log p(c) \tag{3.1}$$

Where, *p(c)* represents the probability of the concept (LCS). The probability *p(c)* was determined by the frequency of occurrence of the LCS and its sub-concepts (the set of concepts subsumed by the LCS) in a corpus. The drawback of this measure is that many concepts share the same LCS in WordNet which leads to assigning the same similarity rating to all the concepts.

Some modifications have been performed to improve the pure information content measure of the original work of Resnik. Jiang and Conrath (1997) presented a hybrid method on the basis of the edge-based notion through adding the information content as a decision factor. If the compared concepts share a lot of information, then the IC of LCS will be high and the semantic distance between the compared concepts and the LCS will be smaller. The proposed measure defined the semantic similarity as the opposite of the semantic distance whereby the concepts with smaller distance are considered more similar to each other than the concepts with a larger semantic distance.

The same elements of the Jiang and Conrath method are used by Lin (1998b) to calculate semantic similarity but in a different fashion. Lin proposed a new formula derived from information theory, which combined information content of the compared words and assuming their independence. The semantic similarity was based on the notion that if the compared concepts share information, then the score of the semantic similarity will be greater otherwise the score of the semantic similarity will be lower.

The majority of subsequent research in the field of the word similarity is either derivative from or influenced by the reported word similarity measures. Liu et al. (2007) proposed an algorithm to calculate the word semantic similarity on the basis of the edge-based notion. This measure used the same elements in (Li et al. 2003),

discussed earlier, but in different fashion which combined the shortest path between the compared concepts and depth of LCS nonlinearly. The fundamental idea of this algorithm was based on the notion that the process of the human judgement for semantic similarity can be simulated via the ratio of common attributes to the total attributes between the compared words. Meng et al. (2014) combined path and information content of concepts to calculate the word semantic similarity. The proposed algorithm used Lin's measure to calculate the information content of concepts. The similarity of the compared concepts is inversely proportional to the path length therefore the proposed algorithm used a nonlinear function to meet this requirement. The overall semantic similarity score was identified by the combination of the Lin's measure with the shortest path of the compared concepts nonlinearly.

3.3 Short Text Semantic Similarity Measures

The current state of short text semantic similarity measures can be categorized into three groups: Corpus based measures, Knowledge based measures and Hybrid measures.

3.3.1 Corpus-based Measures

Corpus-based measures principally use the frequency of a word's occurrence to compute the similarity between short texts. Generally these methods derive the statistical information from the corpus to produce a score of similarity. A well-known early method of this kind is the Latent Semantic Analysis (LSA) (Landauer et al., 1998) which was presented as an information retrieval technique. A set of terms and documents were used to generate a high dimensional matrix which was decomposed by singular value decomposition into three other matrices. To compare two short texts using LSA, two vectors containing the semantic meaning of their words were formed in a reduced dimensionality space and then the overall similarity was calculated by the cosine of the angle between their corresponding row vectors. The drawback of this measure is that the similarity was calculated without using any syntactic information from the compared texts. Consequently for example, the

sentences “The dog hunted the man” and “The man hunted the dog” will be considered as identical.

Islam and Inkpen (2008) proposed another corpus-based method that calculated the text similarity as a combination of three similarity functions (semantic word similarity, string similarity and common word order similarity). Pointwise mutual information (using the British National Corpus (Brown Corpus, 2005)) was employed for measuring the corpus based word similarity. Longest common subsequence string matching was used as a string similarity method to identify any misspelled word in the short texts. Finally, a common word order similarity was employed to incorporate syntactic information in their proposed measure.

3.3.2 Knowledge-based Measures

Knowledge base methods typically use the semantic information derived from a dictionary, thesaurus or ontology for measuring the similarity between short texts. Kennedy & Szpakowitz (2008) used Roget’s thesaurus with a cosine measure for calculating semantic text relatedness. They presented a method of text representation that endeavours to take advantage of the structure found in Roget’s thesaurus and similar lexical ontologies such as WordNet. The text representation method included mapping the text into weighted concepts which were weighted by two criteria (word frequency and specificity). With this weighting method, cosine similarity was used to define the distance between the short texts.

Ho et al. (2010) presented a method (WSD-STS) for measuring text similarity by transforming an existing corpus based method (STS model Islam & Inkpen (2008)) into knowledge based method. The similarity between short texts was computed by the combination of word semantic similarity and string similarity. The word similarity was calculated based on the comparison of actual meaning through the integration of WSD into the adopted word similarity measure. The result of WSD-STS showed that the knowledge based measure performed better than the corpus based measure, which is a baseline measure.

3.3.3 Hybrid Measures

In hybrid methods, both the corpus based and the knowledge based techniques of word semantic similarity are used for measuring the text similarity. The fundamental model of short text semantic similarity, known as STASIS, was proposed by Li et al (2006). In their proposed measure, a joint word set was dynamically formed through the use of all the distinct words in the compared texts. For each sentence, a semantic vector was obtained by combining semantic information from a structured lexical database (WordNet) with information content from a corpus. STASIS incorporated syntactic information by forming the word order vector for each sentence based on a word sequence and location in a sentence. Semantic similarity and word order similarity were calculated based on two semantic vectors and two word order vectors respectively. The overall similarity was defined as a combination of semantic and word order similarity. Much subsequent research in the field of short text similarity are either derivative from or influenced by STASIS such as (Noah et al., 2007), (Liu et al., 2008), (Achananuparp et al., 2008), (Li et al., 2009), (Osathanunkul, 2014), etc. Evidence has also been published which indicated that this measure was successful used in real-world applications such as conversational agents (O'Shea, K. et al., 2010), (O'Shea, K. et al., 2009) and (O'Shea, K. et al., 2008).

Mihalcea et al. (2006) propose another hybrid method that combines the result of six knowledge based measures and two corpus based measures of the word similarity to derive short text similarity measure. The weakness of this measure is that the similarity of words is calculated by eight different methods, which is not computationally efficient.

Feng et al. (2008) use wordnet (to get lexical taxonomy information) and a Brown corpus-based measure for calculating the text similarity with incorporation of direct relevance (obvious coherence between two words) and indirect relevance information (potential relatedness between two words).

Li et al. (2009) combine semantic information derived from wordnet and a corpus with syntactic information obtained through a shallow parsing process. For each compared text, noun phrases, verb phrases and preposition phrases are extracted using shallow parsing. In their proposed measure, they adopted an existing semantic

vector method proposed by Li et al. (2006) to measure the similarities between different kinds of phrases. The overall similarity is calculated based on the combination of semantic similarities of the three kinds of phrases.

Lee et al. (2014) proposed a new sentence similarity algorithm based on grammatical rule and English WordNet ontology. The proposed algorithm calculated the sentence similarity using syntactic and semantic information derived from the compared sentences. An English syntactic parser designed by (Sleator and Temperley, 1995) was utilized to derive the syntactic information which produced a corresponding syntactic structure containing a set of labelled links that connects pairs of words. The proposed algorithm considered the sentences as a sequence of links and directly extracted the semantic similarity from the same or similar links. The relationships between the compared sentences were represented by means of building a limited size set of grammar matrices. The size of this set was selected as a maximum number of the grammar links produced by the parser. Wu and Palmer (1994) measure was used to calculate the similarity between words that the link contains. The overall sentence similarity was determined from grammar information and the word semantic similarity that the links contain.

It can be observed that the majority of the current STSS measures only focus on the similarity of nouns and ignore other parts of speech (Ho et al., 2010) such as verbs, adverbs and adjectives in the computation of STSS. The primary reason is that, STSS measures utilise word similarity measures to calculate the short text similarity and the majority of the current word measures calculate the semantic similarity of nouns due to the richness of the resources that used to support noun semantic similarity. However, the STSS measures (Li et al., 2009, Ho et al., 2010, Lee et al., 2014) that calculate the semantic similarity based on nouns and verbs used the same word semantic similarity techniques to derive the similarity score for both nouns and verbs. Resnik and Diab (2000) have been reported that the problem of identifying verb similarity is different from noun similarity because the representations of verbs are viewed as holding properties such as sub-categorization restriction and event structure that nouns do not. This implies that using the same computational techniques for verbs as for nouns may not be effective because of their different properties. Also (Pedersen et al., 2005) acknowledged that information content and

path based measures are much more effective for identifying the similarity score of nouns while they struggled when including them in a verb experiment.

Also the majority of current short text measures rely largely on computing the similarity between the words in both short texts but does not take the context in which they occur into account and thus affects the final short text similarity score.

3.4 Evaluation of the Semantic Similarity Measures

The only way to identify the quality of a computational semantic similarity measure with confidence is by means of an investigation of its performance compared with human perception (Resnik, 1995, Gurevych and Niederlich, 2005, O'Shea et al., 2013). This requires the use of a benchmark dataset with similarity ratings collected from human participants.

The design process of a word or short text dataset faces three challenges. Firstly, selection of a sample of the word or short text pairs that represents the properties of the language for which the dataset is created. Secondly, collection of similarity ratings that precisely represented the human perception of similarity using a representative sample of participants. Thirdly, determination of the appropriate statistical measures that can be applied to make judgments about the word or short text similarity measures (O'Shea et al., 2013). This section will review the current state of word and short text datasets based on the methods used to meet the three issues of the dataset design process.

3.4.1 Word Semantic Similarity Benchmark Datasets

This section will review the details of the current state of word similarity datasets.

3.4.1.1 R&G-65

Rubinstein and Goodenough (R&G) (1965) produced the most influential word (noun) benchmark dataset for English. A set of 48 English nouns represented in two lists (each list contained 24 nouns) was employed to produce 65 noun pairs. However this dataset was published without justification for the specific choices of 48 nouns and the method used to make up of the combination of 65 noun pairs.

The sample of participants used in the R&G experiment for the collection of human ratings consisted of two groups of college undergraduates with a total of 51 participants. No information was provided as regards the composition of age or gender for each group and whether the sample of participants used in this experiment contained only native English speakers.

A card sorting technique was used for collecting human ratings. A paper questionnaire was used in this dataset to record the results and each of the 65 noun pairs was printed on a separate slip. The order of the 65 slips was randomized before presentation. The participants were asked to sort the slips into order of similarity of meaning to obtain ratings based on "how similar in meaning one word was to another". Each noun pair was rated by assigning a value from 4.0 "near synonymous" to 0.0 "completely unrelated": "the greater the similarity of meaning the higher the number" (R&G 1965).

The semantic similarity score for each noun pair was computed as the mean of the similarity ratings made by the participants. The R&G dataset has been widely used in many experiments for the evaluation of different methodologies using the Pearson correlation coefficient as a measure of agreement. This dataset has indicated stability over the years, where re-rating experiments were carried out with new groups of participants 25 and 30 years later by Miller & Charles (M&C) (1991) and Resnik (1995) respectively. This stability shows that the use of human ratings could be a reliable reference for the purpose of comparing with computational methods.

3.4.1.2 M&C-30

Miller & Charles (1991) replicated the R&G experiment, considering only 30 noun pairs from the 65 noun pairs used in R&G dataset to avoid an inherent bias towards low similarity. This dataset consisted of 10 high similarity, 10 medium similarity and 10 low similarity of meaning noun pairs.

A sample of 38 participants was used in the M&C experiment for the collection of human ratings. All were undergraduate students and native English speakers. No information was provided as regards the distribution of the participants' age, academic background, educational level and gender.

Paper questionnaires were used in this dataset for collecting human ratings. All the noun pairs used in this dataset were printed on two sheets. The order of the 30 noun pairs on the two sheets was randomized before presentation. The participants were asked to examine each of the 30 noun pairs closely and ranked each pair based on 5-point scales which run from 0 "no similarity" to 4 "perfect synonymy".

The semantic similarity score for each noun pair was computed as the mean of the similarity ratings made by the participants. The results of this experiment were reported using Pearson correlation coefficients. The correlation between human ratings in the two datasets (M&C and R&G) obtained a high value of 0.97.

3.4.1.3 Resnik-30

The M&C experiment was replicated by Resnik (1995) in order to obtain a baseline from human ratings for the purpose of comparison. This dataset collected human ratings for the subset of 30 noun pairs used in M&C experiment.

A sample of 10 computer science graduate students and post-doctoral researchers was used to collect human ratings. No information was provided as regards the distribution of the participants' age or gender and whether the sample of participants used in this experiment contained only native English speakers.

The human ratings were collected in this dataset in accordance with the same instruction used by (M&C 1991). However, an electronic version questionnaire of the M&C-30 dataset was used in this experiment and the participants were asked to complete the questionnaire (by mail) in a single uninterrupted sitting.

The semantic similarity score for each noun pair was computed as the mean of the similarity ratings made by the participants. The results of this experiment were reported using Pearson correlation coefficients. This experiment obtained a high value correlation of 0.96 with M&C-30 dataset. The correlation value of 0.96 was considered as a baseline from human ratings and represented an upper bound for the expected performance from a machine computational attempt to carry out the same task.

3.4.1.4 WordSim-353

This dataset of 353 noun pairs was produced with human ratings in 2002 by (Finkelstein et al., 2002). This dataset was published without justification for the method used to generate the set of 353 noun pairs. The set of 353 noun pairs contained the set of 30 noun pairs used in M&C-30 dataset.

A set of 16 non-native English speakers was used to collect human ratings and no information was given about the participants' age, gender, academic background and level and whether the sample of participants used was student or non-student.

No information was provided about the method used in collecting ratings (whether it used online ratings system or paper questionnaire). Also no information was given about randomizing the presentation of the 353 noun pairs. The participants were asked to "estimate the relatedness of the words in pairs". They ranked each pair based on 10-point scales which run from 0 "totally unrelated words" to 10 "very much related or identical words". The semantic similarity score for each noun pair was computed as the mean of the similarity ratings made by the participants.

This review demonstrates that all the reported word datasets were published without justification for the method used to generate the sets of word pairs that were used in the experiments for collecting of human ratings.

3.4.2 Short text Semantic Similarity Benchmark Datasets

There are five short text datasets produced for English which will be reviewed in this section. These are Lee50 (Lee et al., 2005), STSS-65 (Li et al., 2006), Mitchell400 (Mitchell and Lapata, 2008), S2012-T6 (Agirre et al., 2012) and STSS-131 (O'Shea et al., 2013).

3.4.2.1 Lee50

This dataset of 1,225 text pairs was produced with human ratings in 2005 by (Lee et al., 2005). 50 emails of headline stories were collected from Australian Broadcasting news mail service to make a combination of 1,225 unique text pairs. Each text varied in length which ranges from 51 to 126 words.

A sample of 83 University students was used to obtain an average 10 ratings for each text pair. The sample consisted of 29 males and 54 females with the average age of 19.7 years. Each participant was paid with a gift voucher of ten Australian dollars.

Each of the text pairs was presented side by side and between eight to twelve times. The order of left –right position of the texts in a pair and the order of the text pairs' presentation were randomized. The participants were asked to rate each pair based on "how similar they felt the documents were". They ranked each pair based on 5-point scales which run from "highly unrelated" to "highly related". The method used in the selection of the point scale was unspecified.

The results of this experiment were reported as the correlation coefficients but without specifying which type. The average of the correlation of all participants was calculated (which equals to 0.605) and this can be used to assess the performance of a computational method attempt to carry out the same task.

3.4.2.2 STSS-65

This dataset of 65 sentence pairs was produced with human ratings in 2006 by (Li et al., 2006). The sentence pairs were generated by replacing the set of 65 word pairs from (Rubenstein and Goodenough, 1965) with their dictionary definitions from the Collin Cobuild Dictionary (Sinclair, 2001). Each sentence in this dataset varied in length which ranging from 5 to 33 words.

A sample of 32 participants was used in the experiment of the collection of human ratings. All were native English speakers at graduate level or above and they volunteered without compensation. This dataset took a good care to control the distribution of the participants' age, academic background and gender. Regarding to the degree of screening (remove specific participants from the experiment sample), this dataset used the first 32 questionnaires that were returned by participants.

A paper questionnaire was used in this dataset whereby each sentence pair was printed on a separate sheet. The order of the sentences within a pair and the order of 65 sheets within the questionnaire were randomized before presentation. The

participants were asked to rate each pair of sentences based on "how similar they are in meaning". They ranked each pair based on a 5-point scale described as semantic anchors (adopted from (Charles, 2000)) which run from "minimum similarity" to "maximum similarity". Semantic anchors were used as a guide to describe the major similarity scale points used by participants to rank the sentence pairs. This dataset encourage participants assigning a specific degree of similarity by means of use of the first decimal place.

The semantic similarity score for each sentence pair was computed as the mean of the similarity ratings made by the participants. The results of this experiment were reported using Pearson correlation coefficients. The average of the correlation of all participants was calculated (which equals to 0.825) and this can be used to assess the performance of a computational method attempt to carry out the same task. Since its release, this dataset has been widely used for evaluating and comparing new developments (O'Shea et al., 2013).

3.4.2.3 Mitchell400

This dataset of 400 simple sentence pairs was produced with human ratings in 2008 by (Mitchell and Lapata, 2008). The sentence in each pair was three words in length only, generated using intransitive verb (past tense) extracted from CELEX (Baayen et al., 1993) and combined with its subject noun extracted from the British National corpus. Additional information was combined with the verb and subject to construct a sentence such as articles or pronouns. For example, "the horse ran".

The sentence pairs were separated to three blocks which were rated using three samples of 69, 88 and 91 unpaid volunteers' participants. This dataset used only native English speakers and gave a good care to control the distribution of the participants' age and gender. Regarding to the degree of screening, 14 participants who were non-native speakers were removed and also the response of 30 participants was excluded after discovering anomalies in their judgements.

The Webexp online rating system (Keller et al., 2009) was used to collect human ratings. The order of the sentence pairs' presentation was randomized and one

sentence pair was presented to the participants at a time. The participants were asked to rate each pair based on "how similar two sentences are in meaning". They ranked each pair based on 7-point scales which run from "not very similar" to "very similar". The method used in the selection of the point scale was by clicking a button.

The results of this experiment were reported using Spearman's ρ correlation coefficients. The average of the correlation of all participants was calculated which was $\rho = 0.40$.

3.4.2.4 S2012-T6

This dataset of 5,250 text pairs was produced with human ratings in 2012 by (Agirre et al., 2012). This dataset was created as a part of task 6 in SEMEVAL 2012 to train, test and evaluate the algorithms of the semantic text similarity. Each sentence in this dataset varied in length which ranging from 4 to 61 words. The text pairs were constructed using automatic selection methods from several existing corpora which included 1500 sentence pairs which were sampled from the Microsoft Research (MSR) Paraphrase corpus based on 5 bands of string similarity, 1500 sentence pairs were selected from MSR Video Paraphrase corpus based on 4 bands of string similarity, 1500 pairs from Workshops on Statistical Machine Translation (Callison-Burch et al., 2007; Callison-Burch et al., 2008), and 750 sentence pairs from a mapping between the senses of the OntoNotes (Hovy et al., 2006) and WordNet (Fellbaum, 1998).

The Amazon Mechanical Turk online rating system (Buhrmester et al., 2011) was used to crowd source annotations for Human Intelligence Task (HIT). Each HIT contains 5 sentence pairs and this means collecting five annotations per HIT. No information was provided as regards the number of the participants used, the composition of age or gender and whether the sample of participants used in this experiment contained only native English speakers. Each participant was paid $0.20 per HIT. Regarding to the degree of screening, this dataset eliminated participants when their ratings obtained a correlation below 50% with the initial ratings that made by the experimenters on 200 sentence pairs selected randomly from the data.

The Amazon Mechanical Turk online rating system was used to collect human ratings. No information was provided about randomizing the presentation of the sentence pairs. The participants were asked to rate each pair based on "how similar two sentences are to each other". They ranked each pair based on 6-point scales which run from "no different topic" to "completely equivalent as they mean the same thing". Each value in the scale was provided with a definition. The method used in the selection of the point scale was by clicking a button.

The results of this experiment were reported using Pearson correlation coefficients. The Pearson score for each dataset was produced using a simple word overlap algorithm as a baseline to evaluate and compare the performance of the different methodologies.

3.4.2.5 STSS-131

This dataset of 64 sentence pairs was produced with human ratings in 2013 by (O'Shea et al., 2013). The process of the generation of 64 sentence pairs consisted of three steps. First step included selecting a set of 64 stimulus words using a sampling frame technique (Oppenheim, 1992) which is a method of representing a large population with a small carefully-chosen sample randomly selected within constraints. The second step involved producing a database of English sentences using a sample of native English speakers. The set of 64 stimulus words was divided to 4 groups of 16 stimulus words and each participant wrote two sentences for each stimulus word in a specific group. Step three included selecting 64 sentence pairs from the database, which covered varying range of similarity, by three judges. Each sentence in this dataset varied in length which ranging from 5 to 33 words.

A sample of 32 native English speakers was used in the sentence production experiment. All were undergraduate students on Arts and Humanities with a capacity for creative writing. Each participant was paid £5 per hour. Whilst a sample of 64 native English speakers was used in the experiment of the collection of human ratings, consisting of a group of 32 undergraduate students and a group of 32 non-students. Non-student participants volunteered without compensation whilst each student participant was paid £5 per hour. This dataset took good care to control the

distribution of the participants' age, academic background, educational level and gender. Regarding to the degree of screening, 5 participants were removed because they gave ratings to two calibration sentence pairs which differed widely from the ratings provided by 72 participants to the same pairs of sentences in the STSS-65 dataset (O'Shea, 2008).

Human ratings were collected for 64 sentence pairs in accordance with the same procedure used to collect human ratings in STSS-65 dataset.

The results of this experiment were reported using Pearson correlation coefficients. The average of the correlation of all participants was calculated (r equals to 0.891) and this can be used to assess the performance of a computational method attempt to carry out the same task.

It can be observed from the review of the current English datasets that the creation of the dataset involved two important steps: generating the set of short text pairs and collecting human ratings. There is a need for creating a short text dataset (chapter 6 of this thesis) to enable the evaluation of the Arabic short text similarity measure. For the step of generation of the short text pairs, using the automatic selection from corpora as in S2012-T6 (Agirre et al., 2012) can reduce the representativeness (O'Shea et al., 2013). For example, the S2012-T6 dataset was skewed towards the high similarity short text pairs. The use of the dictionary definition as in STSS-65 (Li et al., 2006) narrows the language representation (covering only assertions) (O'Shea et al., 2008). Creation of a short text of the three words in length as in Mitchell400 (Mitchell and Lapata, 2008) is too short particularly as some contain a function word. The method used by STSS-131 dataset (O'Shea et al., 2013) to generate a set of short text pairs will be adapted in this thesis which consisted of selection of a set of stimulus words, asking participants to write short texts using the stimulus word and generation the set of shot text pairs based on human judgements.

For collecting human ratings step, the decision was made to adopt a technique which combined the card sorting with the semantic anchors used in STSS-65 whereby more consistent human ratings (lower noise) was demonstrated by this combination based on the ANOVA experiment on STSS-65 (O'Shea et al., 2010).

3.5 Arabic Resources that Support the Semantic Similarity

Arabic is considered a highly derivational and inflexional language which is spoken and written by more than 300 million people. However, little work has been done on developing linguistic resources for Arabic NLP, especially knowledge rich resources such as ontologies that can support Arabic semantic similarity. Furthermore, only theoretical models are presented and no implementation is available for any of these projects e.g. the work in (Belkredim and El Sebai, 2009) which describes an ontological representation for the Arabic Language. This ontology is relevant because its design is based on Semitic template root-based lexical principles, which represent the Arabic language features but no implementation is available. The Arabic resources used in this thesis will be reviewed in this section based on the availability.

3.5.1 Arabic WordNet (AWN)

AWN is the only freely available lexical resource for modern standard Arabic (Elkateb et al., 2006a). It is based on the design and contents of Princeton WordNet (PWN) for English and can be mapped onto PWN as well as a number of other wordnets. The AWN structure consists of four principal structures. First, the items represent conceptual entities including synonym set (synset), synsets-id (unique identifier), ontology classes and instances. Second, a word entity represents a word sense which contains word form and word-id (used to associate word's citation form with an item). Third, a form entity contains lexical information such as the word's root and broken plural form. Fourth, a link connects in a relation two items such as hyponym, equivalent, etc.

Moreover, the AWN synsets have been mapped to general concepts of an ontology known as Suggested Upper Merged Ontology (SUMO) (Pease and Nile, 2002). SUMO is defined as a language independent ontology which consists of 2000 concepts, 4000 definitional statements and 750 rules (Nile and Pease, 2003). The world is classified by SUMO into upper-level concepts without stating how these general concepts are expressed using terms. An example of these concepts is that the "TransportationDevice" concept. The AWN-SUMO mapping process was performed using three relations which were used to associate the general concepts of SUMO to

the more specific AWN synsets (Elkateb et al., 2006a): synonymy (equivalent links), hypernymy (subsumption links), and instantiation (instance links).

The latest version of AWN consists of 11,270 synsets containing about 23,496 Arabic words which cover nouns, verbs and a very limited number of adjectives and adverbs (AlKhalifa and Rodríguez, 2010). As discussed in Arabic language features section 2.2.1 chapter 2, traditional POS classification incorporates adjectives and adverbs with nouns and there is currently no method to access them in this form. This version of AWN will be utilized in the creation of an Arabic STSS measure in chapter 4 of this thesis which will only focus on nouns and verbs in the short text. Figure 3.1 illustrates a portion of AWN noun hierarchy with SUMO mapping whereby the SUMO general concept *TimePosition* associated to the AWN synset ظهر "noon" by the hypernymy relation.

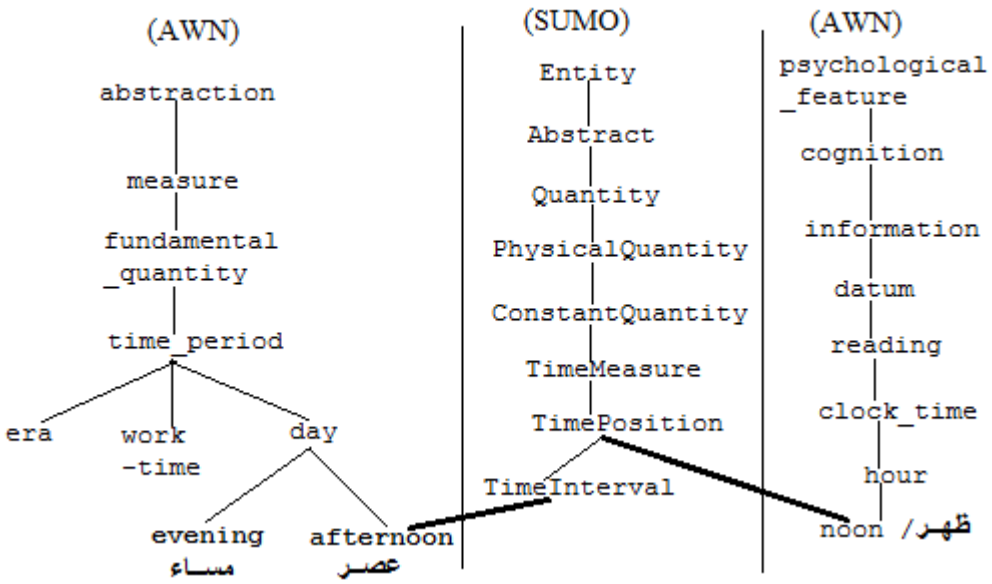


Figure 3.1 Fragment of the AWN with SUMO mapping

3.5.2 Arabic Word Count (AWC)

Attia et al. (2011) produced a large word corpus for Modern Standard Arabic containing one billion Arabic words. This corpus was generated by combining 900 million Arabic words from the Arabic Gigaword corpus (Parker et al., 2009) with 163,649,497 Arabic words collected using news articles from Al-Jazeera website.

This corpus was used to create a large Arabic lexical database of 30,000 lemmas using a machine learning method and a data-driven filtering method. A list of high frequency words for Arabic known as Arabic Word Count (AWC) was created by (Attia et al., 2011) containing the 30,000 lemmata listed according to their frequency with their English glossary and part of speech.

The methodology of the creation of an Arabic short text similarity measure in chapter 4 of this thesis requires weighting each word based on its significance by assigning an information content extract from a corpus. The AWC list will be used to meet this requirement. Moreover, there is a need for materials in Chapter 5 and 6 of this thesis for the creation of data sets to enable the evaluation of both the Arabic verb similarity measure and the Arabic short text similarity measure. This process requires the employment of AWC list. The latest version of AWC list contains 37,700 lemmata which will be used with the work in this thesis.

3.6 Conclusions

This chapter has reviewed the current state of word similarity measures based on an information sources they exploit. It has shown that some of the proposed measures used the information source directly as a metric of word similarity or used a particular information source without considering the contribution of others whilst the best result obtained by the measurement that properly processed and combined the information sources. Also the majority of the current word similarity measures focus on noun semantic similarity. Details of current short text semantic similarity measures have been reviewed which demonstrates that the major challenges faced by existing measures are: understanding context within a short text structure and the use of Part of Speech other than nouns.

Furthermore, the current datasets used in the evaluation process of word and short text similarity measure were reviewed based on the method used to generate the set of word or short text pairs, the sample of participants used, the procedure used in the collection of human ratings and the statistical measures applied to make judgments about the word or short text similarity measures.

Finally, details about the Arabic resources that will be used in chapter 4, 5 and 6 of this thesis were described. The implication of the lack of certain resources used in English will be discussed in chapter 4.

Chapter 4

A Framework for Developing an Arabic Short Text Semantic Similarity Measure

4.1 Introduction

The review of related work in chapter 3 described a number of algorithms which have been developed for measuring Short Text Semantic Similarity (STSS). Most of these are for the English language. To date no STSS measurement has been reported in the literature for Modern Standard Arabic (MSA). This research proposes a novel framework, namely that of NasTa, for developing an Arabic Short Text Semantic Similarity (ASTSS) measure. This in itself requires the following main contributions which include:

- A new Arabic noun semantic similarity (KalTa-A) measure to identify the similarity score between two Arabic nouns.
- A novel Arabic verb semantic similarity (KalTa-F) measure to calculate the similarity between two Arabic verbs.
- A new Arabic word sense disambiguation (AWSAD) algorithm to disambiguate all words (nouns and verbs) in the Arabic short text.
- A novel measurement of Arabic noun and verb Semantic Similarity (KalTa-AF) which is presented to perform word sense disambiguation by calculating the similarity between two words that have a different POS, either a pair comprising a noun and verb or vice-versa.

The development process of the NasTa framework consists of two phases. The first phase relates to the creation of an algorithm, namely that of NasTa-A which is inspired by Li et al.'s algorithm (2006). However, the very rich derivational and inflectional features of Modern Standard Arabic (MSA) mean that the process of creating this measure is not straightforward. The NasTa-A algorithm focuses on noun semantic similarity computation in both short texts which requires creating a new Arabic noun similarity measure to meet this requirement. The second phase of the

development process of the NasTa framework involves developing a new ASTSS algorithm, namely that of the NasTa-F. This algorithm is created to address the weakness of the NasTa-A algorithm which resulted from the properties of the MSA and the drawbacks of the Li et al. measure stated in chapter 3. This requires creating a novel measure for calculating Arabic verb semantic similarity and a new Arabic word sense disambiguation algorithm to disambiguate all words in the Arabic short texts. The two phases of the NasTa framework development process with their requirements are described in this chapter.

4.2 Overview of the NasTa Framework Phase 1

The proposed framework provides a methodology for developing an ASTSS algorithm inspired by Li et al. (2006), namely NasTa-A, which is based on the concepts of semantic nets, corpus statics and word order. The NasTa-A algorithm consists of two fundamental components, the semantic similarity component and the word order similarity component. The computation process of the two components relies on the computation of the word (noun) semantic similarity in both short texts. A search of the literature showed no noun semantic similarity measure has been attempted for MSA. Consequently, a new Arabic Noun Semantic Similarity measure is created to meet this requirement.

The semantic similarity of the two short texts is calculated using information extracted from a structured lexical database known as the Arabic WordNet (AWN) (Elkateb et al., 2006a) and corpus statistics known as the Arabic Word Count (AWC) (Attia et al., 2011). Arabic words exhibit a complex internal structure, as highlighted in chapter 2, whereby a single Arabic word can represent a complete sentence in other languages. An example of this feature is the Arabic word اخبروني (***akbarooni***) which means (they told me). This feature poses an interesting challenge to the STSS computation as the structure prevents the extraction of the semantic information from AWN and AWC directly where the Arabic words have been saved in AWN and AWC as lemmata, as stated in chapter 3. To overcome this challenge, an Arabic morphological analyser is used to obtain the lemma for each word in the input short texts. However, in the lemmatisation process, this challenge impedes the matching of

the word (in the Arabic text) to the correct lemma, resulting in more than one lemma for a given word, each of which may participate in more than one Part Of Speech (POS). Therefore, an Arabic POS tagger is used to address this challenge, the consequence of this is described in section 4.2.1.

The complex internal structure of the Arabic word also requires a method in NasTa-A to represent each word in a short text without losing the specific meanings that are conveyed for a particular context. A joint word set used by Li et al. (2006) is considered suitable for the Arabic short text representation. It is dynamically formed to represent the two short texts based on all their distinct words, for example, the word اخبروني (they told me) and the word اخبرتني (she told me) are considered two different words, the consequence of this is described in section 4.2.3.

Primary syntactical information is incorporated into the NasTa-A algorithm in the form of word order. However, MSA is considered syntactically flexible, i.e. it has a relatively free word order. All the different orders: Subject-Verb -Object (SVO), VSO, VOS are acceptable structures in MSA as described in chapter 2. Therefore it is not possible to extract the corresponding unconstrained Arabic sentence as an English sentence using word order. This algorithm applies only to MSA and the primary word order in MSA is (VSO). Whilst the majority of ordinary modern Arabic speakers use VSO occasionally; occurrences of other order may be observed. Consequently, to investigate the influence of word order in NasTa-A, an Arabic parser presented by (Attia, 2008) is used to manage the syntactical flexibility of MSA by transforming the input short texts to VSO order before submission to the algorithm.

The overall short text semantic similarity is identified by combining the semantic similarity and word order similarity. The framework of the developed measure NasTa-A is shown in Figure 4.1:

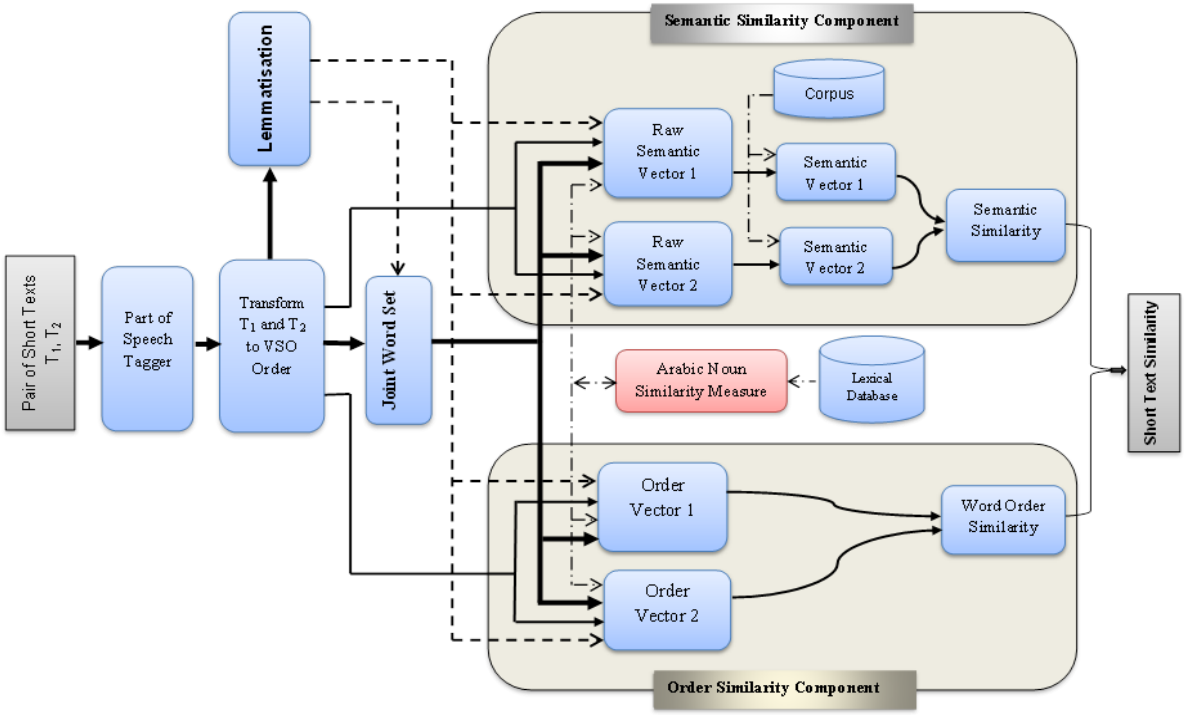


Figure 4.1 Arabic Short Texts Semantic Similarity Framework Phase 1.

A detailed description of each of the NasTa-A components is presented in the following sections.

4.2.1 Arabic Short Text Pre-Processing

The input short texts are pre-processed before their submission to the NasTa-A algorithm which include two steps:

1. **Lemmatisation** – is the task of finding the canonical form, or dictionary form, (which is also named the lemma) for words (Al-Shammari and Lin 2008). For example, the lemma of the Arabic noun **طلاب** (students) is **طالب** (student) whilst the lemma of Arabic word **يعملون** (they work) is **عمل** (worked). The purpose of using lemmatisation is that the words in the AWN and AWC have been saved as lemmata and they are employed by the NasTa-A algorithm to identify the similarity score. The BAMA Arabic morphological analyser (Buckwalter, 2002) is adopted in this research which was identified in chapter 2 as the most suitable because it provides the lemmatised form.

2. **Part of Speech Tagging** – POS tagging is the process of assigning the POS to every word in the short text (Habash, 2010). On account of the complex internal structural feature of Arabic words, BAMA may assign several different lemmata for a given Arabic word each of which may participate in more than one POS. For example, the lemma of the Arabic word كتبهما is either كَتَبَ ***Kataba*** (write) as a verb or كتاب ***Kitab*** (book) as a single noun for the plural كُتُب ***Kotob*** (books). The POS tagger is used to overcome this challenge whereby the lemma of each word in the short text will be selected based on its POS assigned by the tagger. Thus, if the POS assigned by the tagger to the word كتبهما is a verb, then the lemma كَتَبَ ***Kataba*** *write* will be selected.

   In this research, the Stanford POS tagger (Toutanova et al., 2003) for MSA is used to assign the POS to each word in the input short texts which has been stated in chapter 2 as the most suitable because of its accuracy and availability.

Since the noun semantic similarity measure is used to calculate the short text semantic similarity and word order similarity in phase 1, the algorithm for identifying the semantic similarity score between a pair of nouns will be described first.

4.2.2 Arabic Noun Semantic Similarity Measure (KalTa-A)

In this research, a new algorithm namely that of KalTa-A is presented for measuring the semantic similarity between two Arabic nouns. The development of a measurement for calculating the semantic similarity between two Arabic nouns has two requirements.

1. Knowledge resources that support semantic similarity such as ontologies, dictionaries, corpora.
2. An algorithm that utilizes the knowledge resources to identify the word similarity value.

As regards the first requirement, the latest version of AWN described in chapter 3 is the only functional lexical database for MSA which can be used as a knowledge

resource. However, the AWN is a recent development and poses its own interesting challenges when used within applications.

1. As described in section 4.2.1, Arabic words have been saved in AWN as lemmata. The BAMA morphological analyser was used to obtain the lemma for each of the input nouns.
2. Arabic words have been stored with full diacritics in AWN for the purposes of disambiguation. The problem arises because contemporary Arabic words are written without diacritics. For example, the verb "write" has been saved in AWN as كَتَبَ ***kataba*** (with diacritics) whilst in contemporary Arabic writing system it is written as كتب ***ktb*** (without diacritics). The full automatic discretization of the Arabic texts is still in early stages and the most Arabic researchers simply removed the diacritics from the text (Habash, 2010). Consequently, to manage this problem, a de-diacritics process (removing the diacritics from AWN words) is undertaken in order to retrieve words from AWN.
3. Apart from diacritics there are other components of letters which are not handled consistently by humans. Some Arabic letters have the same shape and are only discriminated by adding particular marks which are not diacritics such as a dot, a hamza (ء) or a madda (~) located above or below these letters, as shown in table 4.1. An example of these letters is the Arabic word أداة, whereby the first letter from right to left is *alif* with *hamza* above أ and the last letter *Taa* ة which is *Haa* with two dots above. In contemporary Arabic writing, the Arabic words with these letters are written without marks (hamza, dot and madda) whilst they were stored with marks in the AWN as shown in figure 4.2. In this figure, the word أداة is written without *hamza* above *alif* and without two dots above *Haa* in contemporary Arabic writing.

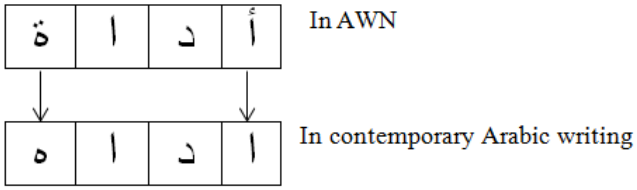


Figure 4.2 the Arabic word اداة in AWN and contemporary Arabic writing.

So these letters are normalised as follows to retrieve from AWN.

1. *Alif* with *madda* آ or *hamza* (أ, إ) are normalized to bare *alif* ا.
2. *Taa* (ة) with *Haa* (ه) without dots.
3. *Alif maqsuura* ى is normalized to "*Ya*" ي

Table 4.1 Arabic letters shared the same shape with different marks.

Letters share same shape	Dot (.)	Hamza (ء)	Madda (~)
Alif ا		أ or إ	آ
Haa ه or ـه	ة or ـة		
Alif maqsuura ى	ي		

Based on the availability of Arabic resources that support semantic similarity, the similarity between the two Arabic nouns is calculated based on a knowledge-based approach. (Hliaoutakis et al., 2006, Pirro, 2009) carried out a comparison between the performances of the reported word similarity measures described in chapter 3. A knowledge based method proposed by (Li et al., 2003) offered the best performance among the reported word similarity measures and has been adopted by many researchers in English. This algorithm is adapted and extended for measuring the similarity between two Arabic nouns.

AWN is constructed in a lexical hierarchy where words are connected with concepts by well-defined types of relations. One simple method for calculating similarity by means of the lexical semantic net is to find the minimum path length that connects the two concepts containing the compared nouns. This is done by finding the meeting point known as the Lowest Common Subsumer (LCS) which is the most specific concept in the hierarchy that subsumes the two concepts, followed by calculating the path distance between them through it. For example, figure 4.3 illustrates a portion of the AWN noun hierarchy. The minimum path length between أب "father" and أم "mother" is 2 (*father* – ***parent*** – *mother*) and the concept والدان "parent" is called LCS for the nouns أب "father" and أم "mother". The minimum path between جد "grandparent" and أب "father" is 6. In this instance, the أم "mother" is more similar to أب "father" than جد "grandparent" to أب "father". If the noun is polysemous then

multiple paths exist between the compared nouns. In this case, the shortest path length between them is used to calculate the similarity.

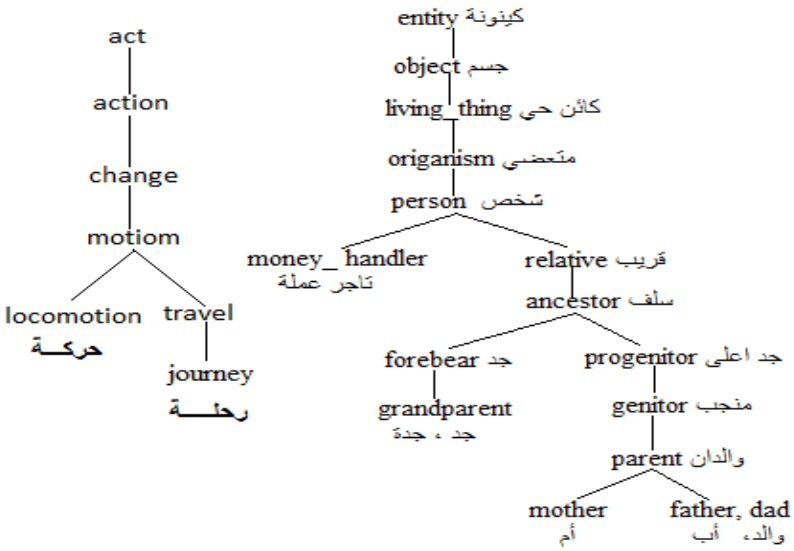


Figure 4.3 A portion of Arabic WordNet noun hierarchy.

Likewise, in figure 4.3, the shortest path length between جد "grandparent" and تاجر عملة" money_handler" is 5, less than from جد "grandparent" to أب "father" which is 6, but it would be incorrect to say that جد "grandparent" is more similar to تاجر عملة "money_handler" than to father. This weakness is addressed by taking the depth of the concept (LCS) in the AWN hierarchy into account in order to adjust the similarity ratings. The depth is calculated by counting the levels from LCS to the top of the noun hierarchy.

Given two nouns n_1 and n_2, the semantic similarity between them as in (Li et al., 2003) can be defined as a function of the attributes of path length and depth as follows:

$$S(n_1, n2) = F(f_1(l), f_2(d)) \quad (4.1)$$

Where, l is the length of the shortest path between n_1 and n_2. d is the depth of the LCS of n_1 and n_2 in a lexical hierarchy. f_1 and f_2 are transfer functions of path and depth respectively.

The similarity interval is [0, 1]. When l =0, the similarity of s (n_1, n_2) = 1 which implies that the similarity is inversely proportional to the path length. Therefore, f_1 is set to be a monotonically decreasing function of l and is selected in exponential form to meet l constraints.

If there is no meeting point between the compared nouns (no LCS), the similarity of s (n_1, n_2) = 0. As shown in figure 4.3, رحلة "journey" and أب "father" are classified under a separate substructure and no LCS subsumes the compared nouns. Hence the similarity between them is 0.

The similarity grows higher if the depth of the LCS of compared nouns increases in a lexical hierarchy which implies that the similarity is directly proportional to the depth. To meet this constraint, f_2 is set to be an increasing function of d.

The overall score of word similarity is calculated by combining the shortest path length and depth of compared nouns nonlinearly using the following formula:

$$Sim(n_1, n_2) = e^{-\alpha l} * \tanh(\beta * d) \quad (4.2)$$

Where, α and β are the length and depth factors respectively, which signify the contribution of the length l, and depth of LCS d. l can be calculated using (4.3):

$$l = d_1 + d_2 - (2*d) \quad (4.3)$$

Where d_1 and d_2 are the depth of n_1 and n_2 respectively. α and β will be calculated in chapter 5.

As a consequence of the nature of the AWN organization scheme, the structure of its hierarchy may produce a bias towards a particular distance computation. As can be observed in figure 4.4, *bus* and *journey* are classified under separate substructures which indicate there is no relationship between them in the AWN hierarchy. This gives a very low machine rating value. However, the definition of *journey* in (Sinclair, 2001) is the act of travelling from one place to another. Whereas the *bus* is a device which serves as the instrument in transportation process which carries the patient of the process from one point to another (Niles and Pease, 2003). In this case,

the KalTa-A measure will be hampered by this weakness as its recall relies on the ontological detail and coverage. However, this weakness can be addressed by means of use of multiple ontologies which provide additional knowledge that may assist in improving the similarity score.

As mentioned in chapter 3, AWN may be augmented with SUMO mappings, which can be exploited to overcome the KalTa-A measure limitation. The SUMO ontology is employed to identify the shortest path length and depth between the compared nouns which are classified under a separate substructure in the AWN hierarchy.

1. Three relations were used to map the AWN synsets to the SUMO concepts, which are synonymy, hypernymy, and instantiation as stated in chapter 3. For example, the noun *journey* in figure 4.4 is associated with the SUMO concept *motion* through the use of the relation hypernym. The KalTa-A measure can benefit from this mapping to augment the relationship between the compared nouns through going across the SUMO hierarchy from the AWN hierarchy to extract the shortest path and depth of compared nouns.
2. The SUMO ontology has a predicate called the *related Internal Concept*. This predicate has two arguments each of which represents a concept in SUMO. It means the two concepts are related within SUMO and there is a significant similarity of meaning between them (Niles and Pease, 2003). The KalTa-A measure can take advantage of this to increase the number of SUMO concepts (by adding the new concepts from the predicate) which are associated with Arabic nouns. This may increase the chances of finding the shortest path length and depth between the compared nouns.

The following example illustrates how the KalTa-A algorithm calculates the similarity between two Arabic nouns. Figure 4.4 illustrates a portion of the AWN noun hierarchy and the mapping to SUMO. To identify the similarity score between the Arabic nouns رحلة *journey* and باص *bus*, the lemma for each noun is obtained using BAMA and the normalization process is performed for each lemma. The shortest path length and depth between the compared nouns is extracted using the AWN noun hierarchy. As shown in figure 4.4, the compared nouns are classified under separate substructures in the AWN hierarchy. This means the similarity score

between them is 0. In this case, the shortest path and depth is calculated using the SUMO ontology. The noun رحلة *journey* is associated with the SUMO concept *motion* whilst باص *bus* is associated with the concept the *TransportationDevice* by the relation hypernym. The *Transportation* concept in figure 4.4 is a related internal concept to the *Transportation Device*. Therefore, the noun باص *bus* is associated with the *Transportation* concept by the relation hypernym which increased the number of the associated concepts to 2. The shortest path and depth of the compared nouns is calculated using SUMO. The shortest path is 4 and the depth of LCS (*motion* concept) is 4. A medium similarity score is obtained by the KalTa-A measure of the compared nouns رحلة *journey* and باص *bus*.

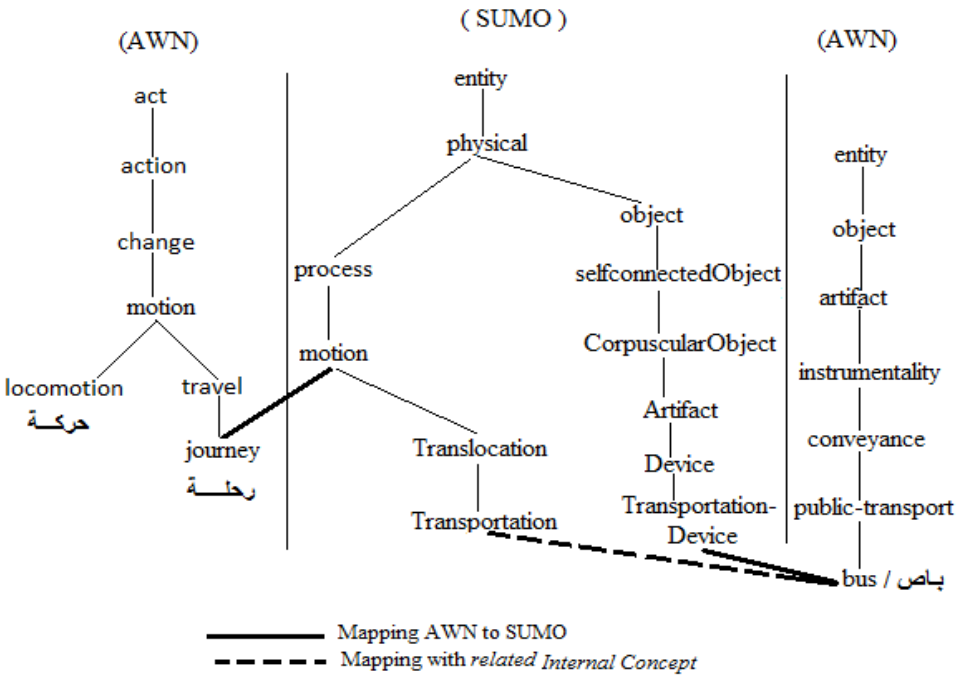


Figure 4.4 Fragment of the AWN with SUMO mapping

The Kal-Ta-A measure should be validated before its integration into the NasTa-A algorithm. The only way to evaluate such a measure meaningfully is by comparison with human perception (Resnik, 1999). Unlike English, Arabic does not yet have a benchmark noun similarity dataset therefore there is a need for a dataset which can be used to identify the quality of the computational Arabic noun semantic similarity algorithms. A substantial experimental methodology was required to create the first noun benchmark dataset for MSA. The methodology used to create this dataset with the procedure for evaluating the KalTa-A measure are presented in chapter 5.

4.2.3 Construction of the Joint Word Set

An important step in calculating the semantic similarity between two short texts is the manner in which they are represented. A short text is made up of a sequence of words. Arabic words exhibit a complex internal structure, whereby a single Arabic word can represent a complete sentence in other languages. For instance, the subject and object of a verb may be embedded within itself. An example of this is the Arabic word اخبروني ***akbarooni*** which means (they told me) whilst the word اخبرني ***akbarani*** means (he told me). With this feature, the NasTa-A algorithm requires a method to represent each word in a short text without missing the specific meanings that are conveyed for a specific context. The solution is to represent the Arabic short texts using all their distinct words (no stemming /lemmatisation). In this example, the word اخبروني (they told me) and the word اخبرني (he told me) are considered two different words. Given two short texts T_1 and T_2, a joint word set T is formed to represent them using all the distinct words in the two short texts from right to left as shown in formula 4.4.

$$T = T_1 \cup T_2 = \{w_1\ w_2 \quad \cdots \quad w_m\} \tag{4.4}$$

For example:

T_1 = اضافة ملعقة عسل الى الحليب كل يوم تعطي طاقة للاطفال

Adding a spoonful of honey to the milk every day gives the children energy.

T_2 = يتناول اولادي الكعك اضافة الى شرب الحليب كل صباح

In addition to drinking the milk, my sons eat cake every morning

As shown in figure 4.5, the joint word set T created for T_1 and T_2 is:

{اضافة, ملعقة, عسل, الى, الحليب, كل, يوم, تعطي, طاقة, للاطفال, يتناول, اولادي, الكعك, اضافة, شرب, صباح}

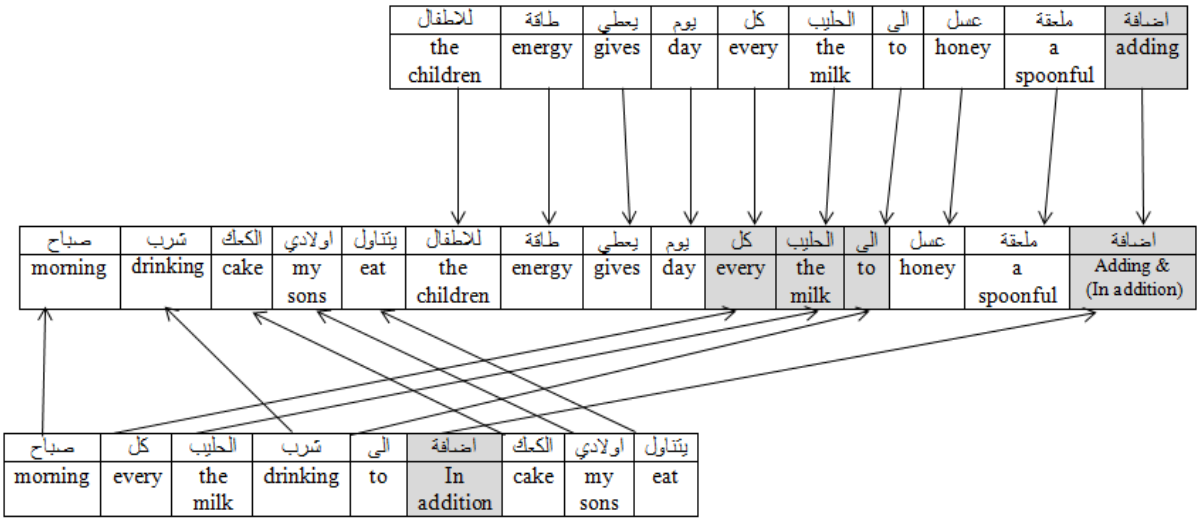


Figure 4.5 joint word set created for the short texts T_1 and T_2.

4.2.4 Semantic Similarity Component

The computation process of semantic similarity between the two short texts is illustrated in this section as follows:

4.2.4.1 Formation of the Lexical Semantic Vectors

For each short text, a semantic vector š is derived from the joint word set. The dimensionality of the lexical semantic vector is equivalent to the number of words in the joint word set, $š_i$ *(i=1, 2, …. , m)* . Each entry value of the lexical semantic vector represents the semantic similarity between the corresponding word in the joint word set and a word in the short text. Equation 4.5 is used to derive the semantic vectors.

$$š = \left(max\left(x_{1,1}, \dots, x_{n,1}\right), max\left(x_{1,2}, \dots, x_{n,2}\right), \dots\dots, max\left(x_{1,m}, \dots, x_{n,m}\right)\right) \qquad (4.5)$$

Where *n* represents the number of words in the short text, *m* represents the number of words in the joint word set. *x* represents the similarity value between a word in the joint word set and a word in the short text. The semantic similarity between the two words is calculated using the KalTa-A measure.

The lexical semantic vector for each short text (T_n) is formed by taking one of the following actions for each word w_i in the joint word set.

- Case 1: If w_i appears in T_n, the entry value of $š_i$ is set 1.
- Case 2: If w_i is not in T_n but w_i and any associated word in T_n have the same lemma, the entry value of $š_i$ is set 1.
- Case 3: Otherwise, the semantic similarity score is calculated between w_i and each word in T_n, using the KalTa-A measure described in section 4.2.2.
  The highest similarity score ς between w_i and the most similar word in T_n is used to set the entry value of š. If ς exceeds a pre-set threshold then $š_i$ = ς, otherwise $š_i$ = 0. If the highest similarity score is below the threshold value, thus the w_i has no meaningful similarity with the word in T_n. In this case, the algorithm uses the threshold to eliminate the noise.

Each word is weighted based on its significance and contribution to the meaning of the short text by assigning an information content extracted from a corpus. The AWC corpus is employed in this research to extract the information content using the following formula:

$$I(w) = 1 - \frac{\log(n+1)}{\log(N+1)} \tag{4.6}$$

Where N is the number of the words in the AWC corpus and n is the frequency of occurrence of the word w in the corpus.

Consequently, each entry value of the semantic vector s_i is weighted according to the information content of w_i (a word in the joint word set) and $\hat{w}_i$ (the associated word in the short text that have the highest similarity score with w_i). Finally, each entry value of the semantic vector s_i is calculated using the formula 4.7.

$$s_i = š \; . \; I(w_i) \; . \; I(\hat{w}_i) \tag{4.7}$$

Where $I(w_i)$ and $I(\hat{w}_i)$ are the information content of a word in the joint word set and its associated word in the short text respectively.

4.2.4.2 Computation of the Semantic Similarity component

Finally, the semantic short text similarity is calculated using the cosine coefficient measure between two semantic vectors s_1 and s_2, as shown in formula 4.8 used by (Li et al., 2006).

$$S_s = \frac{s_1 \cdot s_2}{\|s_1\| \cdot \|s_2\|} \quad (4.8)$$

4.2.5 Word Order Similarity Component

The word order similarity computation process is described in this section.

4.2.5.1 Formation of the Word Order Vectors

The order of the words in two short texts is considered to play an important role in the similarity of meaning of the two texts. The following example illustrates the importance of the word order in the computation of short text semantic similarity.

Example 1:
T_1: القط ركض وراء الفأر / The cat ran after the mouse
T_2: ركض الفأر وراء القط / The mouse ran after the cat

It can be seen from T_1 and T_2 that these sentences contain the same words and are only similar to some extent but clearly very different from the viewpoints of the cat and the mouse. The difference in the word order between T_1 and T_2 results in dissimilarity. Any measure which calculates STSS based on the bag of words approach without taking the position into account considers them to be identical in meaning. Consequently, syntactical information is incorporated into the NasTa-A algorithm in the form of word order. However, Arabic is considered syntactically flexible and has a relatively free word order. All different orders: Subject-Verb-Object (SVO), VSO, VOS are acceptable structures of MSA. In the above example, T_1 (The cat ran after the mouse) can be written as:

1. ركض القط وراء الفأر / (VSO) (Ran-the cat- after the mouse)
2. القط ركض وراء الفأر / (SVO) (The cat ran after the mouse)
3. ركض وراء الفأر القط / (VOS) (Ran-after the mouse- the cat)

These are three valid sentences which have the same meaning in a different word order. This challenge can make calculating the word order similarity (a word sequence and location) much harder to resolve than it is in English.

An Arabic parser presented by (Attia, 2008) is employed to manage the syntactical flexibility challenge. As described in chapter 2, this parser was built within the framework of Lexical Functional Grammar (LFG) throughout the use of the formalisms, tools and common inventory of the Parallel Grammar (ParGram) Group. This parser is available online at http://iness.uib.no/iness/xle-web and allows inputting Arabic sentences and giving the functional-structure (f-structure) as output.

The author (Attia, 2008) claimed that "the challenge of Arabic sentence word order flexibility will melt away in the f-structure, where the Arabic sentence analysis is no different from English or a French one". This is illustrated by taking the following sentence as an example:

ساعدت الحكومة ضحايا الزلزال / the government helped the earthquake victims

This sentence in the VSO order (helped –the government- the earth quake victims) was taken as input by the Arabic parser (using the XLE-web), as shown in figure 4.6 and given the f-structure as output as shown in figure 4.7a.

The same sentence was written in the SVO order الحكومة ساعدت ضحايا الزلزال / (the government helped the earthquake victims). The Arabic parser gave the same f-structure as shown in figure 4.7b.

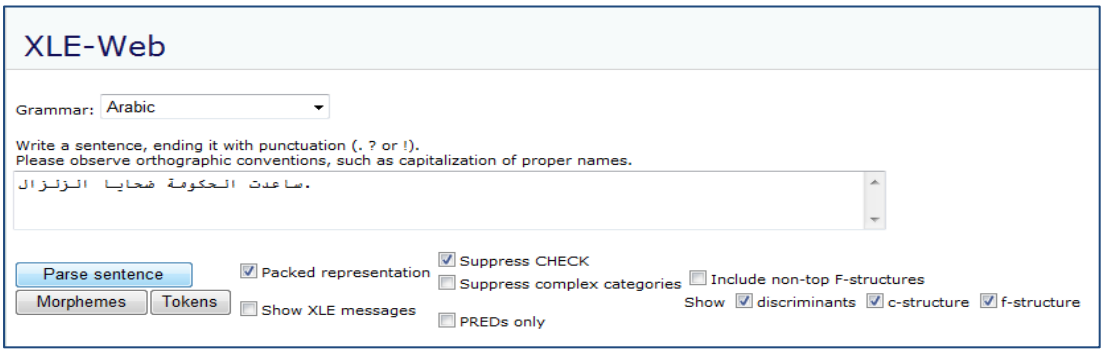
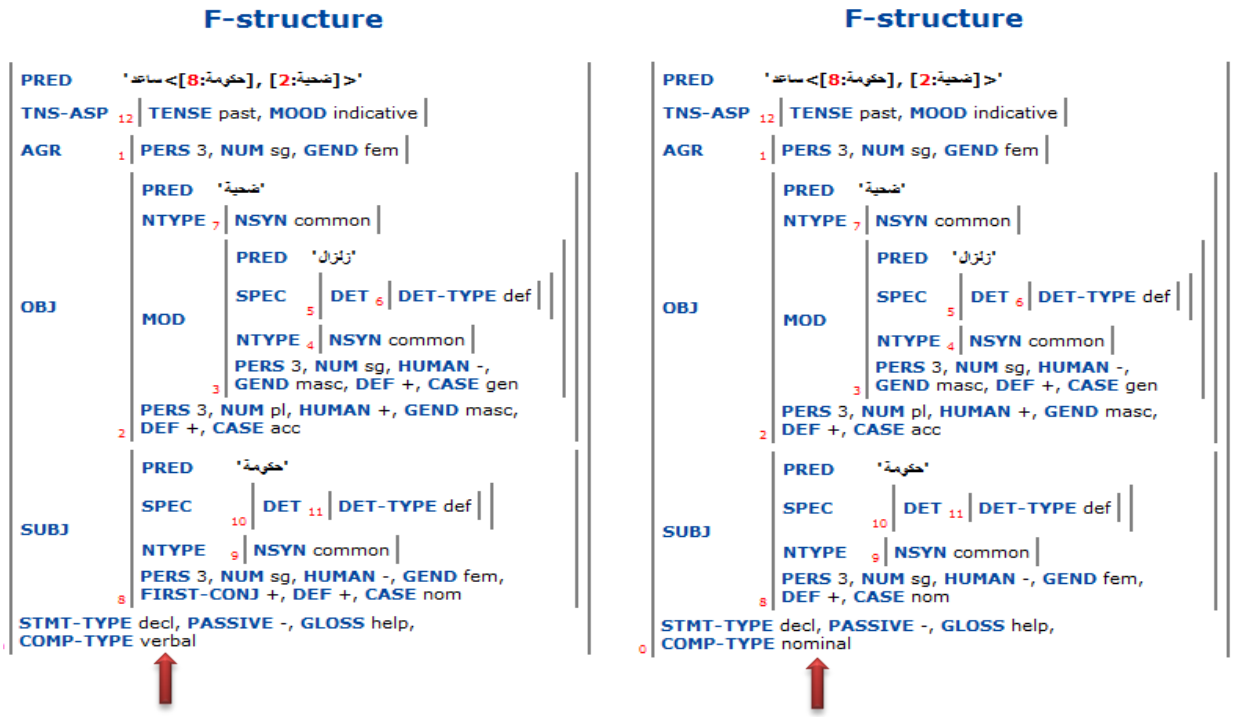


Figure 4.6 XLE-Webs



(a) **(b)**

Figure 4.7 a. F-structure of VSO sentence. b. F-structure of SVO sentence.

It can be observed (figure 4.7 a and b) that, the parser gave the same f-structure for the sentence in the SVO and VSO orders.

The preferred word order in MSA is VSO ((Suleiman, 1989) and (Fargaly and Shaalan, 2009)). Consequently, to address the challenge of a relatively free word

order and to investigate the influence of word order similarity in the NasTa-A algorithm, the input short texts are transformed to the VSO order before submission to the algorithm using the Arabic Rule Based parser. Where the f-structure of each short text is produced using the Arabic parser web page and it is then used to rewrite the short text in the VSO order through the use of the rule applied to build the f-structure itself. This is currently performed manually for research purposes but the output of the parser is suitably structured and tagged to allow this to be automated in the future work.

In the XLE platform, the method of rewriting the sentence is called the generator and is considered the inverse of the parser (Attia, 2008). The generator was used in the translation process where the f-structure for the source language was taken as input and produced the surface string for the target language as output (Attia, 2008). In the NasTa-A algorithm, the surface string is generated for Arabic.

The f-structure in figure 4.7b (sentence in SVO order) is used to generate the same sentence in VSO order as follows:

Rule: *PREDICATE* (0, ساعدت), *SUBJECT* (0, PRED 8), *OBJECT* (0, PRED 2)
PRED (8, الحكومة)
PRED (2, ضحايا), MOD (2, PRED 3)
PRED (3, الزلزال)

Returning to the example in section 4.2.5.1 of two different sentences composed from the same words, the word order similarity of NasTa-A is calculated as follows:

T_1: القط ركض وراء الفأر / *the cat ran after the mouse*
T_2: ركض الفأر وراء القط / *ran –the mouse- after the cat*

In this example, T_1 has SVO order and should transform to VSO order. T_1 f-structure is:

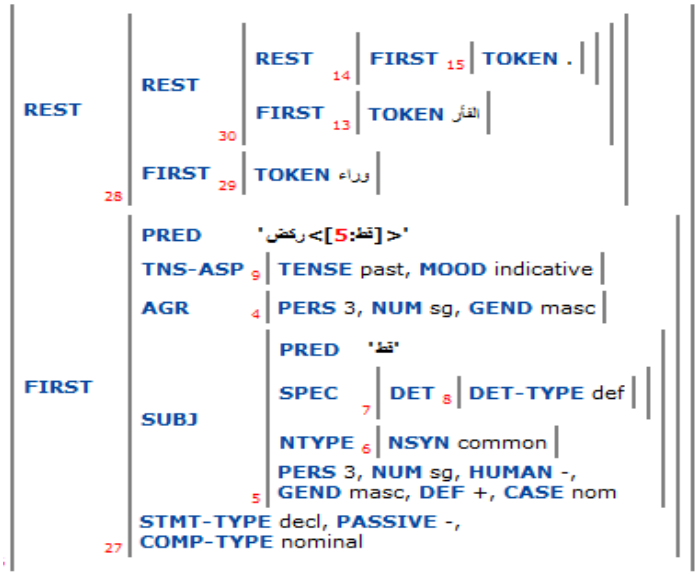


Figure 4.8 F-Structure of T_1

T_1 is transformed to the VSO order using the f-structure and the rule used to build it. The rule is *VERB –SUBJECT –ADVERB PREPOSITION – OBJECT* (Attia, 2008).

After transforming to VSO, T_1 is: ركض القط وراء الفأر / *ran –the cat- after the mouse* and T_2 already has a VSO order ركض الفأر وراء القط / *ran –the mouse- after the cat*

The joint word set T created for T_1 and T_2 is:
T = {ركض, القط, وراء, الفأر}

A unique index number is assigned for each word in the two short texts which is the order that the word appears in the short text. Using the joint word set, word order vectors are produced for T_1 and T_2. These are r_1 and r_2 respectively. For example, r_1 is formed by finding the same or most similar word for each word w_i in the joint word set with the words in T_1. The word order vector r_1 is formed by taking one of the following actions for each word w_i in the joint word set *T*.

1. For each short text T_n
2. For each word w_i in joint word set *T*
3. If w_i appears in T_n, the entry value of *r* is set to the index number of w_i in T_n.

4. If w_i has the same lemma with any associated word in T_n, then the entry value of r is set to the index number of w_i in T_n.
5. Otherwise a semantic similarity is calculated between w_i and each word in T_n to determine the most similar word $\hat{w}_i$ with the highest similarity score ς using KalTa-A measure.

   ς if exceeds a pre-set threshold then the entry value of r is set to the index number of $\hat{w}_i$ in T_n, else it is set to 0.
6. End loop

For this example, the word order vector r_1 is produced for T1 and r_2 produced for T2 using the joint word set T = {ركض, القط, وراء, الفأر}.

$r_1 = \{1, 2, 3, 4\}$ and $r_2 = \{1, 4, 3, 2\}$

4.2.5.2 Calculation of the Word Order Similarity component

Finally, the word order similarity is calculated taking into consideration the number of shared words, their order, the distance between them and the overall length of the short text as shown in formula 4.9 used by (Li et al., 2006).

$$S_r = 1 - \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \quad (4.9)$$

The overall ASTSS is calculated by combining the semantic similarity between two Arabic short texts and Arabic word order similarity as shown in formula 4.10:

$$S(T_1, T_2) = \delta S_s + (1 - \delta) S_r \quad (4.10)$$

Where $\delta \leq 1$ and is used to adjust the relative contributions of semantic and word order information to the final NasTa-A calculation. A complete worked example illustrates how to calculate the two components and the overall short text semantic similarity is given in chapter 6.

The next step of the work must be the evaluation of the NasTa-A algorithm. The only way to identify the quality of a computational STSS measure with confidence is by means of an investigation of its performance compared with human perception (Resnik, 1999, Gurevych and Niederlich, 2005, O'Shea et al., 2013). This will require the use of a STSS benchmark dataset with similarity ratings collected from human participants. No STSS benchmark dataset had been reported in the literature for MSA. Consequently, the first STSS benchmark dataset for MSA and the substantial experimental methodology used for its creation, the procedure for evaluating the NasTa-A algorithm performance and the full experimental results are presented in chapter 6.

4.3 Overview of the NasTa Framework Phase 2

Phase 2 of this research provides a methodology for developing a new ASTSS measure, namely the NasTa-F which is based on the concepts of POS, Arabic Word Sense Disambiguation (WSD) and semantic similarity. The NasTa-F consists of two fundamental components, the Arabic WSD component and the semantic similarity component. The developed measure is created to address the weakness of the NasTa-A algorithm which resulted from the properties of the MSA and the drawbacks of the Li measure (Li et al., 2006) described in chapter 3.

The NasTa-A algorithm focuses only on the similarity of nouns and ignores other Parts of Speech (POS) such as verbs, adverbs and adjectives in the computation of STSS. For example, the same piece of Arabic text ذهب may be a verb "go" or a noun "gold". The NasTa-A algorithm considers these to be the same word throughout the construction of the joint word set. This gives a high similarity between the occurrences in the two short texts which has an impact on the short text similarity score. This drawback is addressed in the development process of the NasTa-F algorithm by calculating the semantic similarity of two short texts based on POS. In the computation process of the semantic similarity component in phase 1, exact lexical matches between words are treated as identical in similarity and the similarity is set to one. For pair of nouns the similarity is computed using AWN as described in section 4.2.2. If a word is a verb, an adjective or adverb, it is treated as its

corresponding noun. In phase 2, exact lexical matches must be from the same POS for identical similarity. Similarities between pairs of nouns are calculated using KalTa-A measure, similar to phase one. There is no verb similarity measure for MSA reported in the literature therefore a novel algorithm is presented in this phase to calculate the Semantic Similarity between pairs of Arabic Verbs (KalTa-F). This measure calculates the similarity based on the assumption that words sharing a common root usually have a related meaning (Rodríguez et al., 2008), which is a central characteristic of MSA. Finally, adjective and adverb pairs either have exact lexical matches where both from the same POS or are rated as unrelated in meaning (0).

As highlighted in chapter 2, Arabic has a higher degree of ambiguity due to a complexity in the Arabic writing system. The reason is that the absence of short vowel representation in MSA resulted in an increase in homographs (words have the same spelling but different pronunciations, and usually with different meanings). As with English, most Arabic words are polysemous (a word has one spelling and pronunciation but also multiple meanings). Take the Arabic word حدث as an example. This word without context and diacritics offers multiple meanings which can mean حَدَثَ ***hadatha*** "happened", حَدّثَ ***haddatha*** "talked" or حَدَثْ ***hadath*** (means juvenile or event). The NasTa-A algorithm relies largely on computing the similarity between the Arabic words in both short texts but does not take the context in which they occur into account and this affects the final short text similarity score. Both homograph and polysemy are instances of the need for WSD which is defined as the process of identifying the correct sense of a particular word based on the context in which it appears (Navigli, 2009). Consequently, the development process of the NasTa-F algorithm addressed this challenge by disambiguating all the words (nouns and verbs) in the input short texts. A new Arabic WSD algorithm is presented which relies on AWN similarity to perform the WSD using three similarity measures. These comprise the KalTa-A measure for calculating the similarity between pairs of nouns, the KalTa-F measure for calculating the similarity between pairs of verbs and a novel measurement of Arabic Noun and Verb Semantic Similarity (KalTa-AF) which is presented to identify the similarity between two words that have a different POS, either a pair comprising a noun and verb or vice-versa.

This section has identified several novel components of an ASTSS framework (NasTa) which may or may not contribute to performance at the current state of art of Arabic NLP. An evaluation of which should actually be incorporated is conducted in chapter 6. The framework of the developed measure NasTa-F is shown in Figure 4.9:

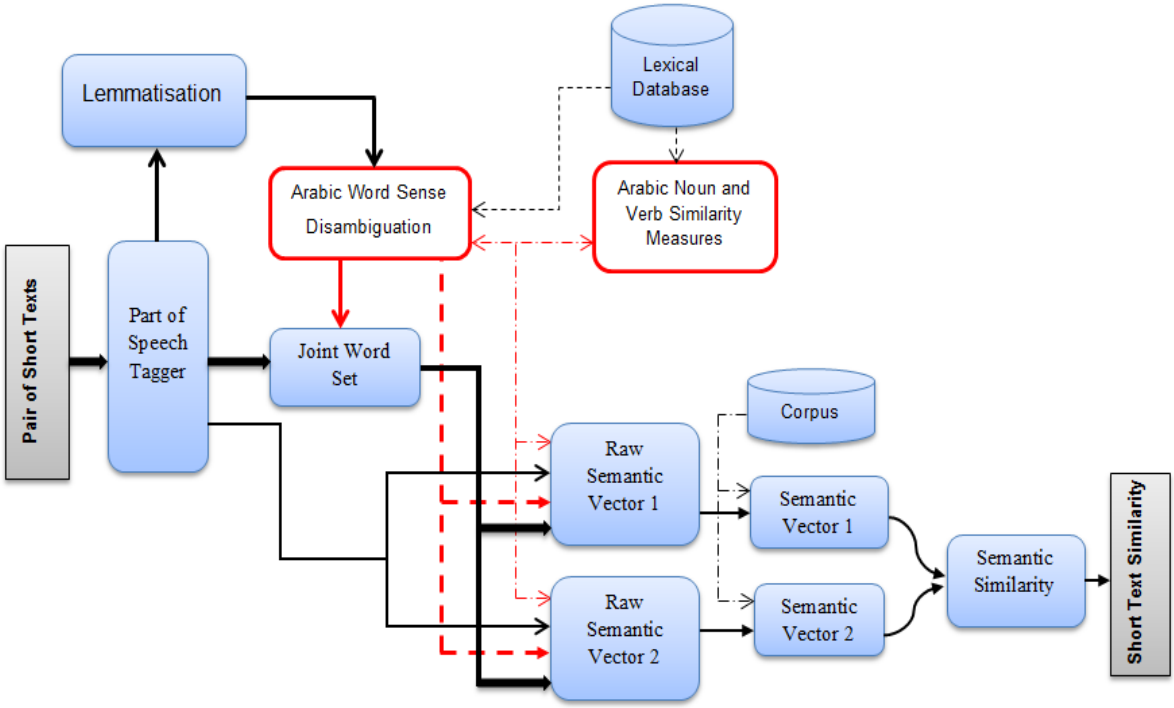


Figure 4.9 Arabic Short Texts Semantic Similarity Framework Phase 2

A detailed description of each of the NasTa-F components is given in the following sections.

4.3.1 Arabic Short Text Pre-Processing

The Stanford POS tagger is used to assign the POS to every word in the input short texts and the BAMA Arabic morphological analyser is utilized to obtain the lemma for each word in the two short texts.

Since the semantic similarity measures (KalTa-A and KalTa-F) are used in performing Arabic WSD and calculating the semantic similarity, it is appropriate for these algorithms to be described first. The KalTa-A measure has already been presented in the first phase (section 4.2.2) whilst the KalTa-F algorithm is described in the following section.

4.3.2 Arabic Verb Semantic Similarity Measure (KalTa-F)

No prior work has been reported as regards the creation of an Arabic verb semantic similarity measure. In this study, a hybrid approach is presented to identify the similarity rating between two Arabic verbs based on the MSA characteristic and the concept of noun semantic similarity. AWN is used as a knowledge resource which supports the semantic similarity of the Arabic verbs. The first step in the methodology of the creation of the KalTa-F measure is to explore the suitability of the Arabic noun semantic similarity algorithm (KalTa-A) for measuring the similarity of words through expanding it to identify the verb similarity scores as follows:

4.3.2.1 KalTa-F Measure:

Given two verbs v_1 and v_2 and using the verb hierarchy in AWN, the shortest path length and the depth of LCS between the compared verbs should be calculated to identify the verb similarity score. However, the verb hierarchy in the taxonomy of AWN is considerably shallower than the noun hierarchy. The nouns in AWN were classified into only 9 noun hierarchies, and they have a tendency to be very deep whilst the verbs were classified into hundreds of hierarchies, and most of these hierarchies are only a few concepts deep. The shallow verb hierarchy in the taxonomy of AWN severely limits the KalTa-F measure effectiveness. Whereby, it is difficult to determine relationships (path length and depth) between verbs that can be used to identify the verb similarity rating using the KalTa-A algorithm directly. Figure 4.10 illustrates a portion of the verb hierarchy in the taxonomy of the AWN.

To illustrate the limitations of the verb hierarchy consider the following examples:

The similarity score of the verb pairs حسب ***Hasaba*** "compute" and عـد ***Ead~a*** "count" was calculated by applying the KalTa-A algorithm for verbs. Using the verb hierarchy, 9 senses were determined for the verb حسب ***Hasaba*** "compute" some of which are shown in figure 4.10 whilst 2 senses were determined for the verb عد ***Ead~a*** "count". The shortest path length between the two verbs was 0. Based on the

path length constraint, the similarity is inversely proportional to path length. The KalTa-F algorithm should give a high similarity score of the compared verbs حسب compute and عـد count. However, a medium similarity rating value was obtained by the KalTa-F algorithm. The reason for this result was that, the KalTa-F algorithm defined the similarity score as a function of the attributes of path length and depth relating to formula (4.2) and (4.3). Due to the shallow verb hierarchy, the depth of the compared verbs (compute and count) from LCS to the top of verb hierarchy was 2. As stated in section (4.2.2), the similarity is directly proportional to the depth, thus a medium machine similarity score between the compared verbs was obtained.

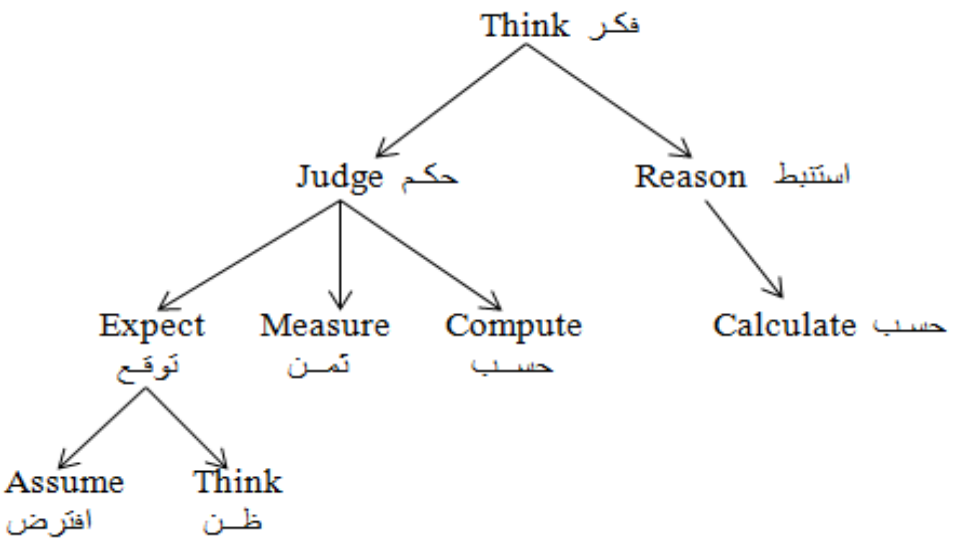


Figure 4.10. A portion of AWN verb hierarchy containing *Hasaba* (compute)

For the same reason, a medium machine similarity rating was obtained for the verb pair قرأ read and تعلم learn. The shortest path between them was 0 but the depth of LCS was equal to 1.

The similarity ratings obtained by KalTa-F measure for the verb pair جاء ***jaA'a*** "come" and وصل ***waSala*** "arrive" presented another example of the verb hierarchy limitation. There was a very low similarity value (*equal to 0*), which indicated that the compared verbs were classified under separate substructures and there was no relationship (shortest path and depth) between them in the AWN verb hierarchy.

4.3.2.2 KalTa-F Final

In spite of the challenge of the verb hierarchy limitations and the sparseness, a novel method is presented to enrich the verb hierarchy based on the assumption that words sharing a common root usually have a related meaning (Rodríguez et al., 2008). This supports the use of path measures between pair of nouns related to the verbs, which have greater richness in the complexity and links for enriching the verb hierarchy.

As previously highlighted in chapter 2, Arabic words within a specific semantic field are generated based on derivation from a root making them related in meaning and form, and assigning their syntactic categories in accordance with particular patterns. On account of this, an Arabic verb is formed by replacing the root in a template, thus guaranteeing a semantic relationship with other verbs that have the same root (McCarthy, 1981). This assumption is employed for enriching the relationships of the compared verbs. Unfortunately, only theoretical models are presented in Arabic ontologies which have been designed based on this assumption and no real-word implementation was available (Belkridem and El Sebai, 2009).

However, the AWN provides lexical (dictionary) information such as the Arabic root for each of the Arabic words in the AWN. In addition, the AWN offers a relation that crosses parts of speech boundaries. This relation connects between the derived forms of noun and verb concepts. Consequently, the decision was made to take advantage of this relationship and the lexical information to enrich the verb hierarchy based on the above assumption.

The Arabic root is used to obtain the verbs which are related in meaning to the compared verbs in order to promote the semantic representation (more senses being considered for each of the compared verbs). This may increase the likelihood of finding a relationship (shortest path length and depth) between the compared verbs. Moreover, the related noun forms for each verb sense are obtained using the relationship that connects between the verbs and nouns as a derivationally related in AWN, as shown in figure 4.11which illustrates the root, related verbs in meaning and related nouns in meaning for the verb حسب ***Hasaba*** ”compute” in the AWN. The related nouns are intended to increase the accuracy of the semantic similarity of

compared verbs. This is achieved by means of the employment of the noun hierarchy to obtain the shortest path length and depth of the LCS of the nouns related to the compared verbs.

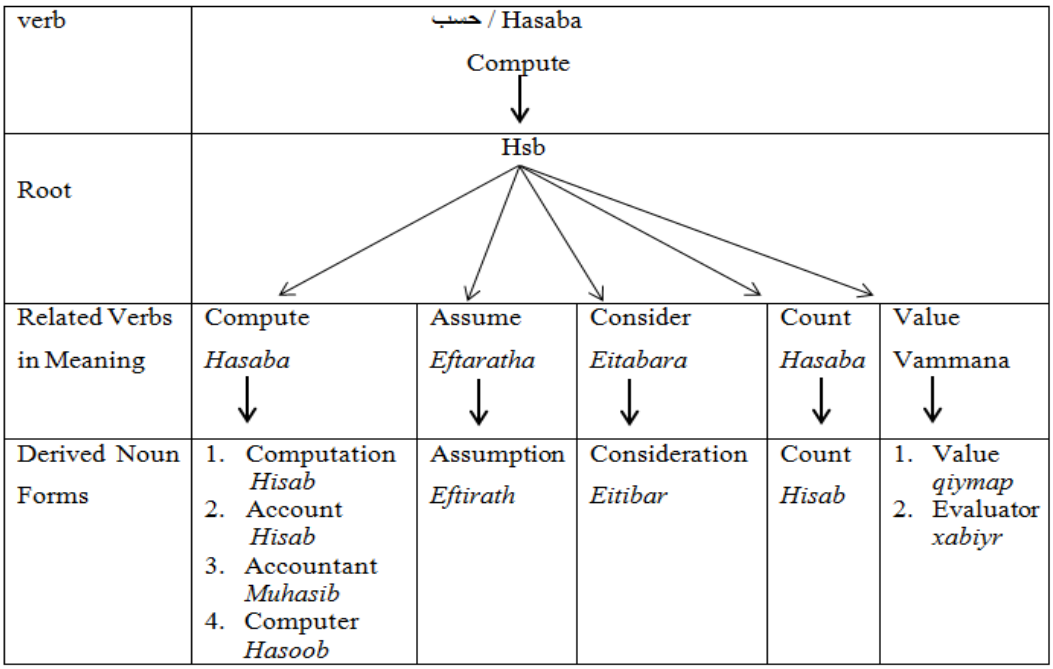


Figure 4.11 The Root, Related Verbs in Meaning and Derived Noun Forms for the Verb *Hasaba* حسب "*Compute*" in AWN.

Given two verbs (V_1, V_2), the score of the semantic similarity between them is identified as follows:

1. For each of the given verbs V_t do, where $t=1, 2$
2. Denote all possible senses of V_t by $\{v_1, v_2, \ldots..v_n\}$.
3. For each verb sense v_k do // $1 \leq k \leq n$, *n is the number of verb senses.*
4. Determine the Arabic root for the sense v_k and denote it as r_k.
5. Determine all related verbs in meaning for r_k and denote as $\{m_1, m_2, \ldots..m_j\}$.
6. For each related verb m_i do // $1 \leq i \leq j$, *j is the number of related verbs.*
7. Determine the derived noun forms which are derivationally related to m_i.
8. End loop
9. End loop // *of sense* v_k
10. End loop // *of given verbs*
11. Calculate the shortest path length and the depth of LCS between all derived noun forms of V_1 and V_2 using formula 4.3 of the KalTa-A measure.

$$l = d_1 + d_2 - (2*d)$$

12. Calculate the overall semantic similarity score between V_1 and V_2 using the formula 4.2 of KalTa-A measure.

$$Sim(V_1, V_2) = e^{-\alpha l} * \tanh(\beta * d)$$

13. End algorithm

For example, consider the same verb pair ***Hasaba*** "*compute*" (has 9 senses) and ***Ead~a*** "*count*" (has 2 senses) of the KalTa-F measure section (4.3.2.1) for the purpose of comparison. Figures 4.11 and 4.12 illustrate the root, related verbs in meaning and related nouns in meaning for the verb حسب ***Hasaba*** "compute" and the verb عـد ***Ead~a*** "count", respectively, in the AWN.

- The first step is to determine the roots for each sense of the compared verbs. All 9 senses of the verb *compute* have the same root which is حسب ***Hsb***. With regard to the verb *count*, its senses also have the same root which is عدد ***Edd***.
- The related meaning verbs are then determined for each root which were 12 for the root ***Hsb*** and 10 for the root ***Edd***. This implies that the number of senses for the verb *compute* حسب become 12 whilst for the verb *count* عـد it was 10. Figure 4.11 shows some related meaning verbs retrieved for the root حسب ***Hsb*** such as "compute" (حسب ***Hasaba***), "assume" (افترض ***Eftaratha***), "consider" (اعتبر ***Eitabara***), "count" (حسب ***Hasaba***), "value" (ثمـن ***Vammana***), etc. Whilst figure 4.12 shows some related meaning verbs retrieved for the root عدد ***Edd***.

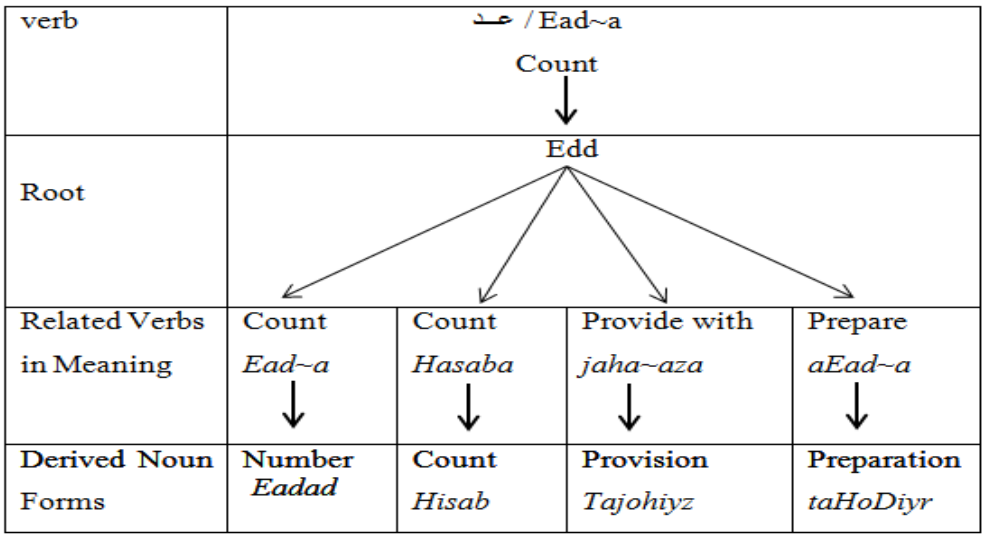


Figure 4.12 The Root, Related Verbs in Meaning and Derived Noun Forms for the Verb *Ead~a* عـد "*Count*" in AWN.

- The related nouns were determined for each of the 12 senses (related meaning verbs) of the verb *compute* حسب. Figure 4.11 shows some related nouns retrieved.

1. For the sense *compute* (حسب ***Hasaba)***, the related nouns are "computation" (حساب ***Hisab***), "account" (حساب ***Hisab***), "accountant" (محاسب ***Muhasib***), "computer" (حاسوب ***Hasoob***).
2. For the sense *assume* (افترض ***Eftaratha***), the related noun is "assumption" (افتراض ***Eftirath***).
3. For the sense *consider* (اعتبر ***Eitabara***) the related noun is "consideration" (اعتبار ***Eitibar***), etc.

Likewise, the related nouns were determined for each sense of the verb *count* as shown in figure 4.12.

1. For the sense *count* (عـد ***Ead~a***), the related noun is "number" (عـدد***Eadad***).
2. For the sense *count* (حسب ***Hasaba***), the related noun is "count" (حساب ***Hisab***).
3. For the sense *provide with* (جهز ***jah~aza***), the related noun is "provision" (تجهيز ***tajohiyz***), etc.

Finally, the shortest path and depth of the compared verbs (*count* and *compute*) was identified using their related nouns. The shortest path value obtained between them equals 0 whilst the depth of the LCS is 9. A high similarity machine rating score of 0.999 was obtained for the compared verbs.

The KalTa-F measure requires validation before its integration into the NasTa-F algorithm. This was done by producing a new Arabic verb benchmark dataset which is the first of its kind for Arabic. The methodology used to create this dataset with procedure for evaluating the KalTa-F measure are presented in chapter 5.

4.3.3 Arabic Word Sense Disambiguation (AWSAD)

The literature survey in chapter 2 distinguished two distinct approaches of the generic WSD, which are the target word and all words in the text. In the target word (or lexical sample) approach, a single ambiguous word is disambiguated in a given context. All words WSD approach includes disambiguating all content word classes

(nouns, verbs, adjectives and adverbs) in a text. This research focuses on all words WSD however a target word approach is highlighted due to its influence on the method presented here.

The literature showed that knowledge-based WSD has become the most promising approach, due to the availability of dictionaries, thesauri, lexical databases and ontologies such as wordnet, which are increasingly enriched (Pedersen et al., 2005).

As stated in section 4.3, ambiguity is considered a big challenge for MSA. Different algorithms of Arabic WSD have been described in chapter 3 but no implementation is freely available in the manner of WordNet::SenseRelate::AllWords for English, and also they are not available from the authors for the purpose of research. In addition, a majority of existing Arabic WSD algorithms were developed to disambiguate a single ambiguous word (target word) in a given context.

In this research, a new algorithm for Arabic WSD namely that of AWSAD is presented to disambiguate all words (nouns and verbs) in the Arabic short texts based on a knowledge-based approach. The AWSAD algorithm performs WSD without requiring any manual training data but uses AWN as a knowledge base. This algorithm utilizes measures of Arabic word semantic similarity to identify the similarity ratings between pairs of nouns, pairs of verbs and noun-verb pairings.

Pedersen et al. (2005) presented a knowledge based algorithm of target words WSD known as the maximum relatedness disambiguation algorithm. This algorithm was described as a general framework algorithm which can be used to perform WSD using any semantic relatedness or similarity measure. The authors investigated several measures of English word similarity as a means of disambiguating a single word in the context.

In this research,

1. Pedersen et al. algorithm is adapted to perform a target word Arabic WSD using three AWN similarity measures developed in this study which are the KalTa-A measure, the KalTa-F measure and the Arabic noun-verb semantic similarity (KalTa-AF) measure (described in section 4.3.3.1).

2. The target word Arabic WSD is extended to disambiguate all words (nouns and verbs) in Arabic short texts.

The proposed algorithm (AWSAD) disambiguates each word in the input short text separately and works from right to left. Each word being disambiguated (known as a target word) is based on its surrounding words which make up its context window. The context window of n size is formed as the target word in the middle and ((n-1)/2) of context words on the left and ((n-1)/2) on the right of the target word. For example, if the window size is 5 there are 2 words on the left of the target word and 2 on the right. However, the number of words on the target word's sides is unequal if the target word appears near to the beginning or end of a short text. For example, if the target word is the end word, there are no words on the left of the target word in the context window. Each target word is disambiguated as follows:

The words in the context window are denoted as $\{w_1, w_2, \ldots, w_n\}$, where the window size is n and w_t is a target word, $1 \leq t \leq n$. Suppose each word w_i has the m_i senses, indicated as $\{s_{i1}, s_{i2}, s_{i3}, \ldots\ldots, s_{im_i}\}$. The AWSAD algorithm intends to disambiguate the target word w_t by assigning one sense from the target word senses $\{s_{t1}, s_{t2}, s_{t3}, \ldots\ldots, s_{im_i}\}$ which has the highest score as the most appropriate sense (intended sense) for w_t. The score of each target sense is identified by comparing it with the senses of its adjacent words in the context using a measure of semantic similarity. For each adjacent word, the algorithm selects the similarity score of the sense that is most similar to the target sense and exceeds the pre-set threshold. The algorithm then adds the score from each of the adjacent words, and this will be the score for the target sense. The target sense with the highest score is assigned as the intended sense for the w_t. The following formula describes in brief the algorithm of disambiguation of the target word, (Pedersen et al., 2005).

$$Correct_{sense} = max_{i=1}^{m_t} \sum_{j=t-c_r, j\neq t}^{t+c_l} max_{k=1}^{m_j} Sim(s_{ti}, s_{jk}) \quad (4.11)$$

Where s_{ti} represents the i^{th} sense of the target word t, and s_{jk} represents the k^{th} sense of the context window word j. c_r represents the number of context word on the right

side of the target word while c_l is the number of words on the left side. Each word in the context window must be known to the AWN otherwise this word is eliminated. $Sim(s_{ti}, s_{jk})$ is the measure of the Arabic word semantic similarity which is used to identify the similarity score between the compared senses s_{ti}, s_{jk}. If the two senses (s_{ti}, s_{jk}) have the same POS (either nouns or verbs), then the KalTa-A and KalTa-F measures developed in this research are used to identify the similarity score between them. If the two senses are from different POS (either (noun-verb) or (verb- noun)), the similarity score between them is calculated using a method (KalTa-AF) of similarity which is described in section 4.3.3.1.

The original KalTa-A algorithm (presented in section 4.2.2) takes two nouns as input and then determines all possible senses of each noun. The next step is to calculate the shortest path length and depth of the LCS of the compared nouns which are used to identify the similarity score between them. In the AWSAD algorithm, the KalTa-A measure used to identify the similarity score between the target word sense and the senses of its adjacent words in the context. This requires modification of the KalTa-A measure to take two noun senses (instead of two nouns) as input and then calculate the path length (instead of the shortest path) and depth of the LCS of the compared senses in order to give the similarity score. Consequently, the modification in formula 4.2 will be only in the definition of *l* (path length).

The KalTa-F algorithm (presented in section 4.3.2) must also be modified to take the two senses of verbs and calculate the similarity between them without determining the verbs' roots as follows:

1. For each of the input sense, determine its derived noun forms only.
2. Calculate the shortest path length and depth of LCS between the derived noun forms of the compared senses.
3. Return the similarity score between the compared senses.

As described earlier, the score of each sense of the target word is calculated by selecting the highest similarity score of each of the surrounding words. The highest similarity score of the surrounding word may be very low which indicates that this word is highly dissimilar. In this case, the algorithm uses the threshold to eliminate

the noise. This provides an element of robustness to polysemy as all possible senses were taken account of.

The following steps illustrate the procedure of disambiguation of all words in an Arabic short text.

1. For all words w_t in the input short text, do // *w_t should be known by AWN, $1 \leq t \leq$ N and N represents the number of words in the short text.*
2. Create *n* - word size window, which includes the target word in the middle, c_r ((n-1)/2) and c_l ((n-1)/2).
   //***c_r is the number of words on the right, c_l is the number of words on the left. c_r = 0, if the target word is the first word and c_l =0, if the target is the end word.***
3. Determine candidate senses s_{im_i} of each word in the window using the AWN.
4. // *disambiguate_target_word*
5. For each sense s_{ti} of the target word w_t, do
6. Set *Sens_score[i]* to 0
7. For each word w_j in the context window, do ($j \neq t$).
8. For each sense s_{jk} of w_j, do
9. Calculate the similarity score *sim-score* between s_{ti} and s_{jk} .
10. End loop.
11. Assign the highest *sim-score* to w_j.
12. If highest *sim-score* > *threshold* then add it to *Sens_score[i]*
13. End loop // *target_word*
14. Choose the $sense_i$ that has the highest score in *Sens_score[i]* as the intended sense to the target word w_t.
15. End loop
16. End procedure

4.3.3.1 The Measurement of Noun-Verb Semantic Similarity

If the Arabic short text has only one verb, the AWSAD algorithm will not be able to disambiguate this verb because there is no other Arabic verb in the context window to compare with and this will limit the AWSAD algorithm effectiveness. The method presented in this section is to address this drawback by means of expanding the AWSAD algorithm for comparison of a pair of senses with a different POS (either

noun-verb pair or verb-noun pair). To illustrate this case, the following sentence provides an example.

ساهم مـدير المدرسـة بـتوزيع الهـدايا على الطـلاب

sAhama	mudiyr Almadrasap	bitawoziyE	AlhadAyA	Ely	AlTulAb
contributed	the headmaster	to distribution of	the presents	to	students

The headmaster contributed to the distribution of presents to students

The AWSAD algorithm will start from right to left and the first target word will be the verb ***sAhama*** ساهم (contributed). The proposed algorithm disambiguates the target word through comparing its senses with the senses of its adjacent words. For this example, all the adjacent words for the verb ساهم are nouns only. To allow for the AWSAD algorithm to compare between a pair of senses with different POS (either verb-noun or noun-verb), a new algorithm namely that of KalTa-AF is presented which takes advantage of the relationship that across the POS in AWN which connects the verbs and nouns as derivationally related. The algorithm takes two senses (verb and noun) as input and returns the similarity score between them as output.

For the input pair verb- noun, the target word is disambiguated as follows:

1. Let us denote all possible senses of verb by $\{v_1, v_2, \ldots..v_l\}$ and all noun senses by $\{n_1, n_2, \ldots\ldots, n_m\}$.
2. For each verb sense v_k do // *$1 \leq k \leq l$, l is the number of the verb senses.*
   - Determine the set of all related noun forms that are derivationally related to v_k using the relationship that connects between the verbs and nouns in the AWN and denote them $R_k = \{r_1, r_2, \ldots\ldots, r_i\}$.
   - For each noun sense n_z do // *$1 \leq z \leq m$, m is the number of the noun senses.*
   - Extract the shortest path and depth between R_k and the noun sense n_z using formula 4.3.
   - Calculate the similarity Sim (v_k, n_z) using formula 4.2.
   - End loop
   - Identify the final similarity for the verb sense v_k

$$v_k = max_{j=1}^{m} Sim(R_k, n_j) \qquad (4.12)$$

Where j represents jth sense of the input noun.

3. End loop
4. End procedure

4.3.4 Construction of the Joint Word Set

As described in section 4.2.3, the joint word set was constructed to represent the Arabic short texts using all their distinct words (no stemming /lemmatisation). However, the joint word set was formed without taking the POS of each word in the short texts into consideration. The NasTa-F is created based on the POS concept therefore the joint word set is formed using all distinct words of the compared short texts and the POS of each words. Consider the same example in section 4.2.3 for the purpose of comparison.

T_1 = اضافة ملعقة عسل الى الحليب كل يوم تعطي طاقة للاطفال

Adding a spoonful of honey to the milk every day gives the children energy.

T_2 = يتناول اولادي الكعك اضافة الى شرب الحليب كل صباح

In addition to drinking the milk, my sons eat cake every morning.

Figure 4.13 illustrates the joint word set created to represent T_1 and T_2.

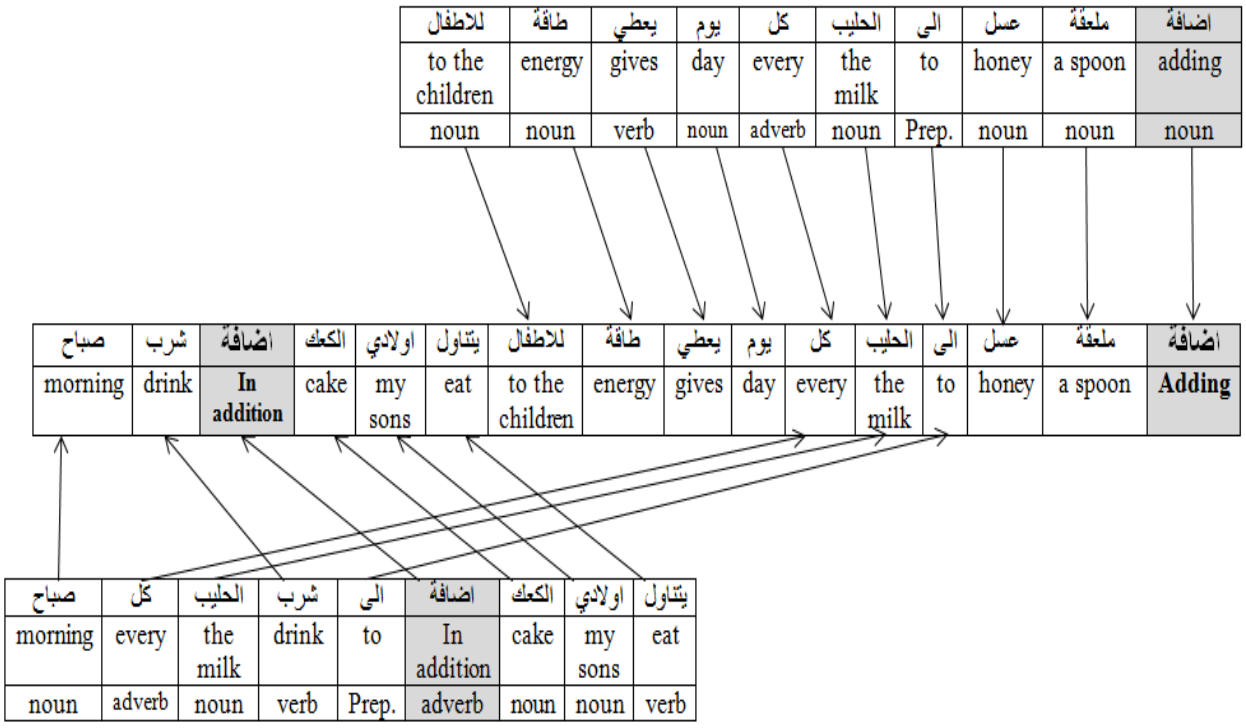


Figure 4.13 joint word set created for the short texts T_1 and T_2.

The two short texts contain the word اضـافة which has the same form but different POS. In T_1 this word appeared as a noun (adding) whilst in T_2 it appeared as an adverb (in addition). This word appeared once in the joint word set created for NasTa-A which formed without consideration of the POS of each word whilst in the joint word set created for NasTa-F; this word appeared twice based on its POS in each short text as shown in figure 4.13. The two short texts contain the word الحليب "the milk" which has the same form and POS therefore it appeared once in the joint word set.

The joint word set is formed as follows:

1. For each short text T_n do
2. For each word w_{ni} in T_n do
3. If w_{ni} not in the joint word set T then add w_{ni} to T.
4. If w_{ni} in T with different POS then add w_{ni} to T.
5. Otherwise, do not add w_{ni} to T
6. End loop
7. End loop

Each word in the two short texts and in the joint word set is paired with the correct sense assigned to this word by the AWSAD algorithm.

4.3.5 Formation of the Lexical Semantic Vectors

For each short text, a semantic vector š is derived from the joint word set. The dimensionality of the lexical semantic vector is equivalent to the number of words in the joint word set, $š_i$ *(i=1, 2, …. , m)* . Each entry value of the lexical semantic vector represents the semantic similarity between the corresponding word in the joint word set and a word in the short text. The semantic vector is derived using formula 4.5.

$$š = \left(max\left(x_{1,1}, \dots, x_{n,1}\right), max\left(x_{1,2}, \dots, x_{n,2}\right), \dots\dots, max\left(x_{1,m}, \dots, x_{n,m}\right)\right) \qquad (4.5)$$

Where *n* represents the number of words in the short text and *m* represents the number of words in the joint word set. *x* represents the similarity value between the

word in joint word set and a word in the short text. The semantic similarity between two words is calculated based on the POS of the compared words using KalTa-A or KalTa-F semantic similarity measures. The two measures take the correct sense assigned by the AWSAD for each of the compared words as input and give the similarity score between them as output.

The lexical semantic vector for each short text is formed by taking one of the following actions for each word w_i in the joint word set.

- Case 1: if w_i appears in T_n and they have the same POS, the entry value of š is set 1. For example, if w_i is the noun اضافة (addition) and the associated word in the short text is also the word اضافة (the same form) but the POS is the adverb (in addition to), then š ≠1. If the two words have the same form and POS such as the verb يذهبون (they go), then š=1.

- Case 2: if w_i has the same lemma and the same POS with any associated word in the short text, then š is set 1. For example, if w_i is the verb يذهبون "they go" which has the lemma ذهب (***Dahaba***) "go" and the associated word is the verb ذهب "go" (the same lemma with w_i ***Dahaba***), then š=1.

- Case 3: for each word in the short text do
  1. If w_i and the associated word have a different form and different lemma but the same POS (noun or verb), then the semantic similarity is calculated between them using the KalTa-A or KalTa-F measure.
     The highest similarity score ς between w_i and the most similar word in the short text is set as entry value of š. Where, if ς exceeds a pre-set threshold then š = ς, otherwise š = 0. If the highest similarity score is below the threshold value, thus the w_i has no meaningful similarity with the associated word.
  2. Otherwise (w_i and the associated word have different POS) the similarity between them is set to 0.

Each word is weighted based on its significance and contribution to the meaning of the short text by assigning an information content extracted from a corpus. An AWC

corpus is employed in this study to extract the information content using the following formula:

$$I(w) = 1 - \frac{\log(n+1)}{\log(N+1)} \quad (4.6)$$

Where *N* is the number of the words in the AWC corpus and *n* is the frequency of the word w in the corpus.

Consequently, each entry value of the semantic vector s_i is weighted according to the information content of w_i (a word in the joint word set) and $\hat{w}_i$ (the associated word in the short text which have the highest similarity score with w_i). Finally, each entry value of the semantic vector s_i is calculated using the formula 4.7.

$$s_i = \check{s} \; . \; I(w_i) \; . \; I(\hat{w}_i) \quad (4.7)$$

Where $I(w_i)$ and $I(\hat{w}_i)$ are the information content of a word in the joint word set and its associated word in the short text respectively.

4.3.6 Computation of the Overall Short Text Semantic Similarity

Finally, the semantic short text similarity is calculated using the cosine coefficient measure between the two semantic vectors s_1 and s_2, as shown in formula 4.8.

$$S = \frac{s_1 \; . \; s_2}{\|s_1\| \; . \; \|s_2\|} \quad (4.8)$$

4.4 Conclusions

This chapter has presented a novel framework for developing an ASTSS measure. The development process of ASTSS framework (NasTa) consisted of two phases. Phase 1 concerned the creation of the NasTa-A measure inspired by Li algorithm which focused on the computation of the noun similarity in both short texts. Further research was needed to extend the NasTa-A measure for understanding context

within a short text structure and the use of POS other than nouns. Consequently, the second phase of the development process of the NasTa framework involved developing a new ASTSS measure (NasTa-F) which covered the POS, Arabic WSD and semantic similarity.

This chapter has also presented four new measures which were used in the computation process of NasTa framework components which included:

- **KalTa-A Measure** – a new algorithm was presented to identify the similarity between pairs of Arabic nouns from a knowledge based approach using information sources extracted from AWN and SUMO. This measure was created to meet the NasTa-A algorithm requirement and used in the computation process of NasTa-A components and NasTa-F components.

- **KalTa-F Measure** – this measure was created in the second phase of NasTa framework development process to meet NasTa-F requirement which calculated the short text similarity based on the POS. A novel algorithm was presented to calculate the similarity between pairs of Arabic verbs based on the assumption that words sharing a common root usually have related meaning which is a central characteristic of MSA. The roots of compared verbs were identified using AWN and employed to determine the related meanings of verbs and nouns of compared verbs. Then, the related nouns were utilized to identify the similarity score of the compared verbs using information sources extracted from AWN.

- **AWSAD Algorithm** – a new Arabic WSD algorithm was presented to disambiguate all words (nouns and verbs) in the Arabic short texts relying on AWN similarity measures developed in this chapter. This algorithm was employed by NasTa-F to address the challenge of missing the short vowel diacritics in the contemporary Arabic writing which causes great ambiguity.

- **KalTa-AF Measure** – A novel algorithm presented to identify the similarity score between two words that have different POS, either pair of noun and verb or pair of verb and noun. This algorithm developed to perform Arabic WSD based

on the concept of noun semantic similarity which takes two senses (either verb and noun or noun and verb) as input and return the similarity score as output.

Some problems were addressed through the development process of NasTa framework which resulted from the properties of the MSA. BAMA morphological analyser and Stanford POS tagger were used to address the challenge of the complex internal structure of Arabic words. Attia's Rule Based parser was used to address the syntactical flexibility of MSA by transforming the short texts to VSO order. AWSAD algorithm was created to address the challenge of ambiguity caused by missing the short vowel diacritics in the contemporary Arabic writing system.

The next phase of the work must be the evaluation of the above new algorithms and determination of which combination should be used profitably in ASTSS framework. This will require the creation of appropriate benchmark datasets and procedures to standardise their use and evaluate the performance of future algorithms developed in this field against those presented here.

Chapter 5

Evaluation of the Arabic Word Semantic Similarity Measures

5.1 Introduction

This chapter describes the evaluation procedures of the Arabic word similarity measures presented in chapter 4: the Arabic noun semantic similarity (KalTa-A) measure and the Arabic verb semantic similarity (KalTa-F) measure. The only way to identify the quality of a computational word similarity measure with confidence is by means of an investigation of its performance compared with human perception (Resnik, 1999, Gurevych and Niederlich, 2005). This requires the use of a word benchmark dataset with similarity ratings collected from human participants.

The first contribution of the work in this chapter is the creation of two Arabic word benchmark datasets: the Arabic noun benchmark dataset and the Arabic verb benchmark dataset. These datasets are the first of their kind for Modern Standard Arabic (MSA). The methodology used for creating these datasets comprises five fundamental steps including the gathering of materials, generation of word (noun or verb) pairs, collection of human ratings, computation of the overall ratings and validation of the datasets.

The created datasets are then used to assess the accuracy of the KalTa-A and KalTa-F measures. The evaluation process involves partitioning each dataset into training and evaluation sets. The training datasets are used to identify the optimal values of KalTa-A and KalTa-F measure parameters, whilst the evaluation datasets are used to assess the accuracy of each measure. The second contribution of the work in this chapter includes the methodology used in the process of partitioning each dataset, the process of the optimization of parameters in the algorithms and the procedure used to assess the accuracy of each measure.

5.2 Creation of an Arabic Noun Benchmark Dataset (ANSS-70)

Creating this dataset, namely that of the ANSS-70, required a substantial and sound experimental method which was divided into three major stages including:

1. Selecting the stimulus nouns
2. Constructing the set of Arabic noun pairs based on human participants.
3. Collecting the human similarity ratings for the set of Arabic noun pairs generated in the second stage.

The literature review in chapter 3 highlighted that R&G (1965) created the first English noun dataset using a set of 48 nouns to make up a combination of 65 noun pairs which spanned the range of semantic similarity from minimum to maximum. This dataset was published without justification for the specific choices of 48 nouns and the method of the combination of noun pairs. Later researchers (Miller and Charles, 1991) and (Risnek, 1995) who replicated the R&G experiment used a subset of 30 noun pairs from the 65 pairs of the R&G dataset to remove bias towards low similarity pairs.

This chapter describes a systematic process to select a set of Arabic stimulus nouns which were then employed to make up a combination of Arabic noun pairs based on human judgements to avoid bias towards low similarity in the R&G dataset. The stages of creation of an ANSS-70 dataset are presented in the following sections.

5.2.1 Selecting the Stimulus Nouns

The major step in the production of the ANSS-70 dataset was that of selection of a set of stimulus nouns which represent the nouns in the Arabic language. This was achieved by carefully choosing 56 stimulus words by means of the employment of categories known as category norms. These categories are important and well known word classes (psychology), independent from WordNet and other ontologies.

As stated in chapter 2, a category norm is a set of words, listed by frequency and generated as responses to a specific theme by human participants. An example of these categories is that of the English category norms as presented by Battig and Montague (1969). Using these categories offers the opportunity of distributing a small sample of noun pairs through semantic space providing better representation of the overall population of noun pairs.

No prior work has been reported on Arabic category norms, hence 27 Arabic categories were produced which cover different semantic themes and contain ordinary Arabic words. The words in each category enjoyed greater similarity to each other than to the words of other categories. The steps of the production of Arabic categories are illustrated as follows:

Step1. R&G used a set of 48 nouns to create the English noun dataset which appear to be 24 pairs of synonyms. These pairs of synonyms appear to be similar, but not identical to the category norms used in Battig & Montague (1969). Therefore, to take advantage of four decades of experience with the R&G dataset, the decision was made to assign these pairs to semantic categories consistent with Arabic nouns. Consequently, twenty-two usable categories were generated from R&G using the following process:

1. For each English pair of nouns (pair of synonyms), the nouns were translated into Arabic using the first meaning from an established English–Arabic dictionary (Baalbaki, 1987). To ensure translation precision, the translated nouns were verified by a professional translator and a lecturer fluent in both languages.

2. Based on the definition of the noun pair (Sinclair, 2001), an Arabic category was given a specific name. A set of Arabic nouns within the same category theme (described in one word) were appended to generate an entire category.

   For example, the pair of synonyms *Gem* and *Jewel* were translated into (جوهرة) in Arabic. The Arabic category was created and named the Gemstones category (احجار كريمة) based on the definitions of jewel (*a precious*

stone used to decorate valuable things that you wear, such as rings or necklaces) and gem (*a jewel or stone that is used in jewellery*). A set of Arabic nouns within the same category theme (Diamond / ماس, Pearl /لؤلؤ, Crystal / بلور ...) were added (using Battig & Montague category members for guidance) to create an entire category. However, some English nouns have been omitted due to translation problems. For example, the noun *madhouse* was translated into the two-word term, "*Mustashfa Almajaneen"* مستشفى المجانين which was therefore omitted. On the other hand, two English nouns were translated into a single Arabic noun as in *Gem* and *Jewel* example. This was added to the category and more examples sought to make up the shortfall.

***Step2*.** In order to promote the semantic representation and incorporate particular Arabic themes, five new categories were created which consisted of ordinary Arabic nouns. For example, the Arabic categories created in the first step have the type of male life stages category, thus to expand this theme and include both males and females, the type of female life stages category was created. Religious events and type of lifestyle categories were produced to incorporate particular Arabic themes. Table 5.1 presents the list of Arabic category names.

27 Arabic categories generated in step 1 and 2 were employed to select a set of stimulus Arabic nouns. This set should be selected and presented by means of a method that contributes to the control of the semantic similarity range (maximum to minimum) covered by the set of Arabic noun pairs which are generated at a second stage. This was achieved by selecting the first two nouns from each category to produce a set of 56 stimulus nouns. This set was represented into two columns of 28 nouns (A and B) to create a List of Arabic Nouns (LAN). Each column contained a noun from each theme such as *Hospital* in column A and *Infirmary* in column B, as shown in table 5.2. The LAN is used in the second stage to generate a set of Arabic noun pairs.

Table 5.1 The List of Arabic Categories Names

	Categories Names	اسماء الفئات العربية
1	Medical Places	مواقع طبية
2	Handwritten text	نص مكتوب يدويا
3	Type of male's life stages	مراحل حياة الذكر
4	Member of the clergy	رجل دين
5	Transportation vehicles	مركبات نقل
6	Coastal area	منطقة ساحلية
7	Bird	طير
8	Type of furnishings	نوع من المفروشات
9	Source of a human body energy	مصدر طاقة جسم الانسان
10	Appliance for cooking	جهاز طهي
11	Gemstones	أحجار كريمة
12	Drinking utensil	ادوات او آنية للشرب
13	Geographic	جغرافية الارض
14	Parts of day	اجزاء اليوم
15	Type of equipment	نوع من معدات/ تجهيزات
16	Type of departure	نوع من رحيل/ مغادرة
17	Somebody practices witchcraft	شخص يمارس السحر
18	Wise person	شخص حكيم
19	Facial expressions	تعابير الوجهة
20	Material for tying things	مادة لربط الاشياء
21	Person in slavery	شخص في العبودية
22	Burial place	اماكن لدفن الاموات
23	Religious events	احداث دينية
24	Type of lifestyle	نوع من نمط / اسلوب الحياة
25	Type of female life stages	مراحل حياة الانثى
26	Vacation activities	انشطة العطلات
27	Family members	أعضاء العائلة

Table 5.2 List of Arabic Nouns (LAN)

Column A			Column B		
1	Hospital	مستشفى	1	Infirmary	مشفى
2	Signature	توقيع	2	Endorsement	تصديق
3	Boy	صبي	3	Lad	فتى
4	Master	سيد	4	Sheikh	شيخ
5	Coach	حافلة	5	Bus	باص
6	Coast	ساحل	6	Shore	شاطئ
7	Stove	موقد	7	Oven	فرن
8	Cushion	مسند	8	Pillow	مخدة
9	Slave	عبد	9	Odalisque	جارية
10	Journey	رحلة	10	Travel (noun)	سفر
11	Gem	جوهرة	11	Diamond	الماس
12	Glass	كأس	12	Tumbler	قدح
13	Forest	غابة	13	Woodland	أحراش
14	Hill	تل	14	Mountain	جبل

15	Noon	ظهر	15	Midday	ظهيرة
16	Tool	اداة	16	Means (noun)	وسيلة
17	Food	طعام	17	Vegetable	خضار
18	Wizard	ساحر	18	Magician	مشعوذ
19	Sage	حكيم	19	Thinker	مفكر
20	Smile	ابتسامة	20	Laugh	ضحك
21	Cord	حبل	21	String	خيط
22	Hen	دجاجة	22	Pigeon	حمامة
23	Sepulcher	ضريح	23	Grave	قبر
24	Feast	عيد	24	Fasting	صيام
25	Countryside	ريف	25	village	قرية
26	Run (noun)	جري	26	Walk (noun)	مشي
27	Brother	أخ	27	Sister	أخت
28	Girl	فتاة	28	Young woman	شابة

5.2.2 Experiment 1: Constructing the Set of Arabic Noun Pairs

One of the fundamental obstacles to the production of the ANSS-70 dataset is being able to select a sample of noun pairs that precisely represents the considerable range of noun pairs which can be generated using the set of stimulus Arabic nouns. Furthermore, to assess the accuracy of computational methods effectively, the set of Arabic noun pairs should be generated spanning the range of similarity of meaning from maximum (identical in meaning) to minimum (unrelated in meaning). Semantic similarity judgements are a matter of human perception. Consequently, an experiment was conducted to construct a representative sample of 70 noun pairs based on human judgements.

The R&G dataset used 48 nouns to make up a combination of 65 noun pairs. Later researchers (Miller and Charles, 1991) and (Risnek, 1995) who replicated the R&G experiment used a subset of 30 noun pairs (30 useable pairs) from the 65 pairs of the R&G dataset to remove bias towards low similarity pairs. In the Arabic noun dataset, a set of 56 stimulus nouns generated in section (5.2.1) was used to create a set of 70 noun pairs. The size of the Arabic dataset of 70 noun pairs was sufficiently accurate to assess the accuracy of Arabic noun similarity algorithms because 70 pairs allowed the use of 30 pairs for testing (equaling R&G), plus 30 pairs for setting parameters. The additional 10 pairs provided a safety margin for issues such as one word in a pair

being missing from a language resource (i.e. Arabic WordNet (AWN)). The procedure of creating the set of Arabic noun pairs is described in this section.

5.2.2.1 Participants

Selecting a representative sample of participants who represent the general human population is another challenge for the process design of the ANSS-70 dataset. The value of a sample of participants selected to carry out a specific experiment could be reduced as a representative sample if there is a great homogeneity of participants (O'Shea, 2010). The sample of the human population used in this study should be representative of native Arabic speakers demographically in terms of their gender, age, education, countries, etc. The reason for controlling the demographics is to prevent confounding factors. As this dataset was produced for the Arabic language, the decision was made to use a sample of 22 native Arabic speakers from different Arabic countries taking into consideration participant academic background, educational level, gender, and age. Previous work (O'Shea, 2010) suggests a minimum size of 16 participants will suffice however more questionnaires were distributed to allow for non-returns. In fact, 22 questionnaires were returned by the deadline and all were used in this experiment.

The participants were from 5 Arabic countries which included: Iraq (7 participants), Jordan (3), Saudi Arabia (6), Libya (3), and Palestine (3). The participants consisted of 10 academics (University lecturers) and 12 non-academics comprising 13 females and 9 males. They were 10 non-students and 12 students. 13 participants were from Science/Engineering backgrounds whilst 9 came from Art/Humanities backgrounds. The participants' educational level included 5 who held bachelor's degrees, 7 who held master's degrees and 10 held PhDs. The average age was that of 34 years with the standard deviation (SD) 6.3.

5.2.2.2 Materials

The list of Arabic nouns LAN (table 5.2) created in section (5.2.1) was presented to the 22 participants for the purposes of generating a set of Arabic noun pairs. The

order of Arabic nouns in column B was randomized to minimize the ordering effects. Each of the 22 Native Arabic speakers was given an envelope containing:

1. Ethics statement
2. a sheet of instructions for producing the noun pairs
3. a LAN sheet
4. two recording sheets to create two lists of Arabic nouns pairs which included:
    - High Similarity of Meaning (HSM) list containing noun pairs between strongly related and identical in meaning.
    - Medium Similarity of Meaning (MSM) list containing noun pairs between vaguely similar and very much alike in meaning.
    - In this experiment the Low Similarity of Meaning (LSM) list was selected randomly resulting in noun pairs which are unrelated in meaning.

5. The final sheet contained minimal details about the participants including name, age, degree and a confirmation that the participant was a native Arabic speaker.

Appendix 1 contains examples of experimental materials including the appendix 1.1 Ethics statement, appendix 1.2 instruction sheet, appendix 1.3 recording sheet and appendix 1.4 personal information sheet.

5.2.2.3 Experimental Procedure

The LAN sheet contains two lists of 28 nouns known as column A and column B. The two lists of 28 nouns were presented to the 22 participants and they were instructed to create a list of 28 HSM noun pairs in order to obtain 23/24 HSM candidate pairs of nouns. The participants were asked to perform the following procedure.

1. Using the LAN sheet, please write a list of 28 HSM noun pairs.
2. Each noun pair must contain one noun from column A and one from column B.
3. The HSM list contains noun pairs between strongly related and identical in meaning.
4. Please write 28 pairs of nouns since all uncompleted questionnaires must be ignored.

The instruction sheet also included notes to enable the participants to create pairs of nouns by selecting any noun more than once from column A with different nouns from B and to avoid rewriting the same pair of nouns on the same sheet or on another sheet.

The same lists of 28 nouns were used to create a set of MSM noun pairs. However, it is relatively difficult for humans to write pairs of nouns of medium similarity between (vaguely similar and very much alike in meaning). Thus, in order to increase the opportunity of obtaining 23/24 MSM candidate noun pairs, the participants were requested to write 32 MSM noun pairs in accordance with the same procedure used to create 28 HSM noun pairs.

5.2.2.4 Experimental Results

The final set of 70 Arabic noun pairs was selected using the HSM and MSM lists generated by participants plus the randomly selected LSM list. Table 5.3 illustrates the final set of Arabic noun pairs, where the first and last columns represent the set of Arabic noun pairs in English and Arabic. The second column contains the number of participants who chose the noun pair. The final set of Arabic noun pairs was selected as follows:

1. 24 noun pairs written by all 22 participants were selected from the HSM list to represent the high similarity of meaning range in the final set of Arabic noun pairs.

2. 23 noun pairs written by more than half the participants were chosen from the MSM list to represent the medium similarity range for the final set of Arabic noun pairs.

3. In order to achieve a good balance in the number of noun pairs in each similarity range, 23 noun pairs were chosen to represent the low similarity of meaning range for the final set of Arabic noun pairs. These noun pairs were selected as a combination of candidate noun pairs chosen as medium similarity by a low number of raters plus low similarity noun pairs selected randomly.

Low similarity noun pairs selected randomly as follows:

For each noun in the LAN, the frequency of appearance of this noun in the final set of Arabic noun pairs was calculated. The nouns that have an occurrence of more than twice were removed from the LAN to avoid a biased set of nouns from being used. The remaining Arabic nouns were used to randomly generate a list of Arabic noun pairs. High and medium similarity noun pairs already found by participants were removed. The remaining pairs were selected at random as they were good candidates for low similarity.

Table 5.3 The Final Set of Arabic Noun Pairs

	Word Pairs		Participants	أزواج الكلمـــات	
High Similarity Noun Pairs					
1	Boy	Lad	22	فتـى	صبـي
2	Coast	Shore	22	شاطئ	ساحـل
3	Cushion	Pillow	22	مخـدة	مسنـد
4	Gem	Diamond	22	الماس	جوهرة
5	Glass	Tumbler	22	قدح	كـأس
6	Forest	Woodland	22	أحراش	غابـة
7	Noon	Midday	22	ظهيرة	ظهـر
8	Tool	Means	22	وسيلـة	اداة
9	Journey	Travel	22	سفـر	رحلـة
10	Smile	Laugh	22	ضحـك	إبتسامة
11	Countryside	Village	22	قريـة	ريـف
12	Girl	Young woman	22	شابـة	فتـاة
13	Signature	Endorsement	22	تصديـق	توقيـع
14	Coach	Bus	22	بـاص	حافلـة
15	Hen	Pigeon	22	حمامة	دجاجة
16	Sepulcher	Grave	22	قبـر	ضريـح
17	Run	Walk	22	مشي	جـري
18	Hospital	Infirmary	22	مشفى	مستشفى
19	Master	Sheikh	22	شيـخ	سيـد
20	Wizard	Magician	22	مشعوذ	ساحـر
21	Feast	Fasting	22	صيـام	عيـد
22	Food	Vegetable	22	خضـار	طعـام
23	Stove	Oven	22	فـرن	موقـد
24	Hill	Mountain	22	جبل	تـل
Medium Similarity Noun Pairs					
25	Sage	Thinker	21	مفكر	حكيم
26	Cord	String	21	خيـط	حبـل
27	Slave	Odalisque	21	جاريـة	عبـد
28	Brother	Sister	21	أخت	أخ
29	Hen	Oven	20	فرن	دجاجـة
30	Coach	Means	19	وسيلـة	حافلـة
31	Sage	Sheikh	18	شيـخ	حكيم

32	Girl	Sister	16	أخت	فتـاة
33	Journey	Shore	15	شاطئ	رحلـة
34	Coast	Mountain	14	جبل	ساحل
35	Master	Thinker	14	مفكر	سيـد
36	Coach	Travel	14	سفر	حافلـة
37	Food	Oven	14	فرن	طعام
38	Brother	Lad	13	فتـى	أخ
39	Girl	Odalisque	13	جاريـة	فتـاة
40	Slave	Lad	13	فتى	عبـد
41	Feast	Laugh	13	ضحك	عيـد
42	Hospital	Grave	12	قبر	مستشفى
43	Hill	Woodland	12	أحراش	تـل
44	Journey	Bus	12	باص	رحلـة
45	Tool	Tumbler	12	قدح	اداة
46	Run	Shore	11	شاطئ	جري
47	Tool	Pillow	11	مخدة	اداة
Low Similarity Noun Pairs					
48	Sepulcher	Sheikh	10	شيخ	ضريـح
49	Cord	Mountain	9	جبـل	حبـل
50	Gem	Young woman	8	شابة	جوهرة
51	Countryside	Vegetable	7	خضار	ريـف
52	Glass	Fasting	6	صيـام	كـأس
53	Forest	Shore	5	شاطئ	غابـة
54	Noon	Fasting	4	صيـام	ظهـر
55	Glass	Diamond	3	الماس	كـأس
56	Signature	String	2	خيط	توقيـع
57	Boy	Midday	1	ظهيرة	صبـي
58	Wizard	Infirmary	0	مشفى	ساحر
59	Cushion	Diamond	0	الماس	مسنـد
60	Noon	String	0	خيـط	ظهـر
61	Boy	Endorsement	0	تصديـق	صبـي
62	Gem	Pillow	0	مخدة	جوهرة
63	Cord	Midday	0	ظهيرة	حبـل
64	Countryside	Laugh	0	ضحك	ريـف
65	Hill	Pigeon	0	حمامة	تـل
66	Slave	Vegetable	0	خضار	عبـد
67	Smile	Village	0	قريـة	إبتسامة
68	Stove	Walk	0	مشي	موقـد
69	Coast	Endorsement	0	تصديـق	ساحـل
70	Smile	Pigeon	0	حمامة	إبتسامة

5.2.3 Experiment 2: Collecting the Human Similarity Ratings

This experiment was conducted to collect human ratings for 70 pairs of nouns generated in experiment 1.

5.2.3.1 Participants

In prior work on word and text semantic similarity various sizes of participant samples were used to collect human ratings. R&G used a sample of 51 undergraduates whilst (Miller and Charles, 1991) used a sample of 38 students. O'Shea (2010) offered evidence that using a sample of 32 participants is sufficient for the collection of good quality ratings, however, they demonstrated that the statistical significance of a sample of participants increased by raising the sample size to 64. In ANSS-70 dataset, the target was to use a sample of 64 participants for the purposes of collecting human ratings but only 60 questionnaires were returned by the deadline and were used in this experiment. The sample of 60 participants was chosen on the basis of its being representative of the general population with equal balance between students and non-students.

1. All were Arabic native speakers who had not taken part in experiment 1 and they were from 7 Arabic countries including Saudi Arabia (16), Iraq (14), Egypt (8), Jordan (7), Libya (7), Palestine (5), and Kuwait (3).

2. The participants' academic backgrounds consisted of 39 Science/Engineering vs. 21 Art/Humanities. Balance was obtained with regard to educational levels and the overall breakdown qualifications were illustrated in table 5.4.

Table 5.4 Participants' educational background

Student	**Non-student (highest qualification)**
11 undergraduate	13 Bachelors
3 Masters	4 Masters
16 PhD	4 PhD
None	9 Diplomas (roughly equivalent to an old UK - BTEC HND).

3. In case of age, the average was 29 years and the standard deviation (SD) was 7.2. Table 5.5 shows the age distributions of a selected sample.

Table 5.5 Age distributions for the Arabic population sample.

Age range	Participants	
18-22	13	Student 11
		Non-student 2
23-29	17	Student 4
		Non-student 13
30-39	23	Student 13
		Non-student 10
40-49	6	Student 2
		Non-student 4
50-59	1	Student 0
		Non-student 1

4. An equal balance was achieved between females and males. The gender balance achieved for non-students was (12 males and 18 females) whilst for students it was (18 males and 12 females).

5.2.3.2 Materials

Each of the 70 noun pairs was printed on a separate card and the cards were presented to the participants for rating how similar the noun pair on each card was in meaning. Each participant was given an envelope containing 70 cards and 3 sheets which included: instructions for collecting the human ratings, a similarity rating recording sheet and a personal information sheet which covered name, age, gender, academic background and confirmation of being a Native Arabic speaker. The 70 cards were randomly ordered before presentation to reduce the ordering effects. Appendix 2 contains examples of experimental materials which include:

- Appendix 2.1 instruction sheet.
- Appendix 2.2 recording sheet
- Appendix 2.3 a sample card.

5.2.3.3 Experimental Procedure

A further challenge of the design process of the Arabic dataset was to collect ratings that precisely represented the human perception of similarity. The decision was made to adopt a technique which combined the card sorting with the semantic anchors (O'Shea, 2010) whereby more consistent human ratings (lower noise) was demonstrated by this combination notably as regards the unsupervised collection of ratings from the general population sample. Semantic anchors describe the major similarity scale points used by participants to rank the noun pairs. Table 5.6 illustrates the semantic anchors for the five scale points used in this experiment.

Table 5.6 Semantic Anchors

Rating Scale	**Semantic Anchor**	
0	The word pairs are unrelated in meaning	زوج الكلمات لا يوجد ارتباط بينها في المعنى
1	The word pairs are vaguely similar in meaning.	زوج الكلمات بينها تشابه ضمني في المعنى
2	The word pairs are very much alike in meaning.	زوج الكلمات التي بينها تشابه واضح (اكثر من ضمني)
3	The word pairs are strongly related in meaning	زوج الكلمات التي بينها علاقة قوية في المعنى
4	The word pairs are identical in meaning	زوج الكلمات المترادفة او المتطابقة في المعنى

The participants were asked to sort the cards into four groups' accordance with the similarity of the meaning. The HSM group contained noun pairs between strongly related and identical in meaning. The High MSM groups contained noun pairs very much alike in meaning, whilst the Low MSM groups contained noun pairs which were vaguely similar in meaning and the LSM contained noun pairs unrelated in meaning. After sorting the cards, the participants were asked to check them carefully and then rank each noun pair using a point on a rating scales described by the semantic anchors which ran from 0.0 (unrelated in meaning) to 4.0 (identical in meaning). The instruction sheet also included some notes which enabled participants assigning an accurate degree of similarity by means of use of the first decimal place and to avoid using values lower than 0.0 or greater than 4.0 to rate the noun pairs.

5.2.3.4 Experimental Results

The human similarity ratings collected in experiment 2 were calculated as the mean of the judgements provided by the 60 Arabic native speakers for each pair of nouns. Table 5.7 represents the results of experiment 2 which contains the set of 70 Arabic noun pairs with human ratings of similarity. The second and last columns represent the set of Arabic noun pairs in Arabic with approximate translation to English. The third column contains the mean of similarity rating collected from 60 Arabic native speakers whilst the fourth column represents the Standard Deviation (SD) of each noun pair which demonstrates an inevitable degree of noise in human ratings.

Table 5.7 The Arabic Noun Benchmark Dataset

	Noun Pairs		Human Ratings	SD	أزواج الكلمـــات	
1	Coast	Endorsement	0.03	0.14	تصديق	ساحـل
2	Noon	String	0.03	0.18	خيط	ظهـر
3	Cushion	Diamond	0.06	0.24	الماس	مسنـد
4	Gem	Pillow	0.07	0.25	مخدة	جوهرة
5	Stove	Walk	0.07	0.25	مشي	موقـد
6	Cord	Midday	0.08	0.27	ظهيرة	حبـل
7	Signature	String	0.08	0.33	خيط	توقيـع
8	Boy	Endorsement	0.12	0.37	تصديـق	صبـي
9	Boy	Midday	0.16	0.39	ظهيرة	صبـي
10	Slave	Vegetable	0.16	0.42	خضار	عبـد
11	Smile	Village	0.18	0.38	قريـة	إبتسامة
12	Smile	Pigeon	0.20	0.39	حمامة	إبتسامة
13	Wizard	Infirmary	0.22	0.41	مشفى	ساحر
14	Noon	Fasting	0.29	0.44	صيـام	ظهـر
15	Hill	Pigeon	0.33	0.54	حمامة	تـل
16	Countryside	Laugh	0.34	0.56	ضحك	ريف
17	Glass	Diamond	0.36	0.60	الماس	كـأس
18	Glass	Fasting	0.38	0.57	صيـام	كـأس
19	Cord	Mountain	0.54	0.68	جبـل	حبـل
20	Hospital	Grave	0.83	0.81	قبر	مستشفى
21	Forest	Shore	0.86	0.77	شاطئ	غابـة
22	Gem	Young woman	0.87	0.87	شابة	جوهرة
23	Sepulcher	Sheikh	0.89	0.77	شيخ	ضريـح
24	Tool	Pillow	0.99	0.98	مخدة	اداة
25	Coast	Mountain	1.06	0.91	جبل	ساحل
26	Run	Shore	1.13	0.82	شاطئ	جري
27	Hill	Woodland	1.19	0.89	أحراش	تـل
28	Countryside	Vegetable	1.24	0.83	خضار	ريـف

29	Tool	Tumbler	1.32	0.95	قدح	اداة
30	Master	Thinker	1.36	0.87	مفكر	سيد
31	Feast	Laugh	1.36	0.84	ضحك	عيد
32	Hen	Oven	1.44	0.84	فرن	دجاجة
33	Journey	Shore	1.47	0.69	شاطئ	رحلة
34	Coach	Travel	1.60	0.70	سفر	حافلة
35	Food	Oven	1.76	0.79	فرن	طعام
36	Slave	Lad	1.77	0.93	فتى	عبد
37	Journey	Bus	1.83	0.72	باص	رحلة
38	Girl	Odalisque	1.96	0.82	جارية	فتاة
39	Feast	Fasting	1.96	0.98	صيام	عيد
40	Coach	Means	2.07	0.90	وسيلة	حافلة
41	Brother	Lad	2.15	0.78	فتى	أخ
42	Sage	Sheikh	2.26	0.92	شيخ	حكيم
43	Girl	Sister	2.38	0.73	أخت	فتاة
44	Hill	Mountain	2.60	0.84	جبل	تل
45	Hen	Pigeon	2.61	0.83	حمامة	دجاجة
46	Master	Sheikh	2.66	1.07	شيخ	سيد
47	Food	Vegetable	2.78	0.70	خضار	طعام
48	Slave	Odalisque	2.84	0.90	جارية	عبد
49	Run	Walk	3.01	0.81	مشي	جري
50	Brother	Sister	3.08	0.62	أخت	أخ
51	Cord	String	3.09	0.78	خيط	حبل
52	Forest	Woodland	3.14	0.62	أحراش	غابة
53	Sage	Thinker	3.30	0.73	مفكر	حكيم
54	Gem	Diamond	3.38	0.66	الماس	جوهرة
55	Cushion	Pillow	3.38	0.64	مخدة	مسند
56	Journey	Travel	3.39	0.71	سفر	رحلة
57	Countryside	Village	3.41	0.71	قرية	ريف
58	Smile	Laugh	3.48	0.58	ضحك	إبتسامة
59	Stove	Oven	3.55	0.69	فرن	موقد
60	Coast	Shore	3.56	0.69	شاطئ	ساحل
61	Signature	Endorsement	3.58	0.71	تصديق	توقيع
62	Tool	Means	3.68	0.52	وسيلة	اداة
63	Noon	Midday	3.70	0.66	ظهيرة	ظهر
64	Boy	Lad	3.71	0.52	فتى	صبي
65	Girl	Young woman	3.74	0.47	شابة	فتاة
66	Sepulcher	Grave	3.75	0.62	قبر	ضريح
67	Wizard	Magician	3.76	0.53	مشعوذ	ساحر
68	Coach	Bus	3.80	0.50	باص	حافلة
69	Glass	Tumbler	3.82	0.38	قدح	كأس
70	Hospital	Infirmary	3.91	0.28	مشفى	مستشفى

5.2.4 Discussion

5.2.4.1 The Arabic Noun Benchmark Dataset (ANSS-70)

The ANSS-70 dataset is intended to evaluate and compare algorithms running on a scale from minimum (zero) to maximum similarity. This is known as a ratio scale, which was used for both word semantic similarity measures and datasets as a measurement scale (R&G, 1965, Miller&Charles, 1991 and Resnik, 1999). The correlation coefficient is considered a suitable statistic that can be applied for measures made on a ratio scale (Blalock, 1979). In this study, the Pearson product moment correlation coefficient was used to identify the consistency of similarity judgments for each participant with the rest of group. This was undertaken using the leave-one-out resampling technique (Resnik, 1995). The correlation coefficient for each of the 60 participants was calculated between the participant's ratings and the average ratings of the rest of group. Figure 5.1 shows the correlation coefficients of 60 participants on the ANSS-70 dataset.

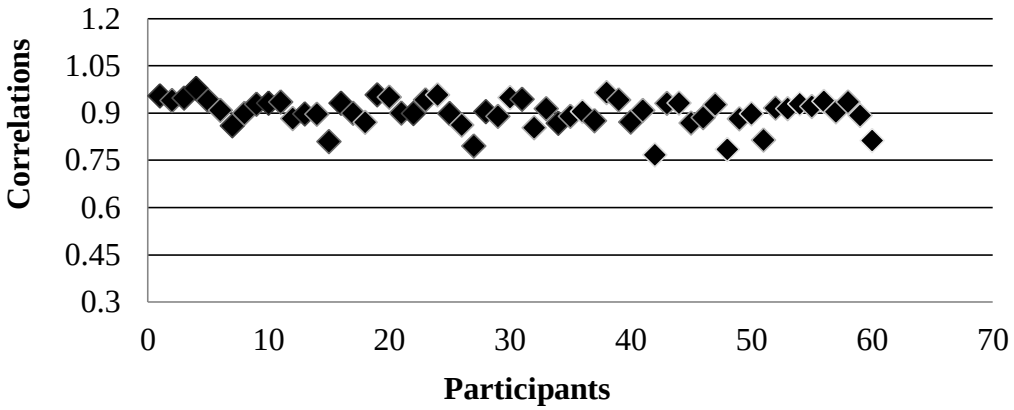


Figure 5.1 The Correlation Coefficients of 60 Arabic Participants

The average of the correlations of all participants on the ANSS-70 dataset was calculated; this can be used to assess the performance of a computational method attempt to carry out the same task. Any noun semantic similarity measure which equals or exceeds the average of the correlations of all participants is considered to be performing well. As shown in table 5.8, the average of the correlations of all participants for the ANSS-70 dataset is 0.902, a good target for a machine algorithm. The worst performing participant of 0.767 is considered as the lower bound for the

expected performance whereas any machine measure coming close to the best performing participant at 0.974 would be considered as performing very well.

Table 5.8 Correlation Coefficient with Mean Human Judgments

	Correlation r
Average of the correlation of all participants	0.902
Best participant	0.974
Worst participant	0.767

Figure 5.2 shows the distribution of the similarity ratings in the full ANSS-70 dataset. The dataset is well balanced, if one considers that ~ 1/3 of the noun pairs are high, ~ 1/3 low and ~ 1/3 across the broad, difficult medium similarity band from 1.0 - 3.0.

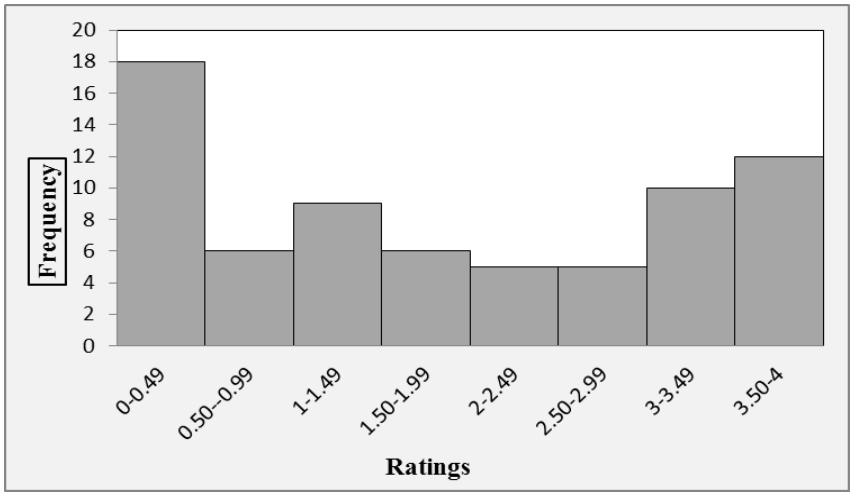


Figure 5.2 Distribution of the similarity ratings in ANSS-70 dataset.

Both high similarity and low similarity noun pairs are subject to very consistent human judgments, as shown in figure 5.3 and figure 5.4. Unlike the low and high similarity noun pairs, the human ratings of the medium similarity noun pairs spread more evenly across the similarity range (0 to 4). Consequently, the medium similarity noun pairs have higher values of SD than the other noun pairs. For example, the noun pair 46 (شيخ سيد) has SD 1.07 and the mean of human ratings 2.66. The distribution of the human ratings for this noun pair should be grouped around a peak of 2.66. In fact the modal class is 3 and the distribution is relatively flat as shown in figure 5.5.

Figure 5.3 Histogram of similarity ratings for noun pair 01, SD= 0.14.

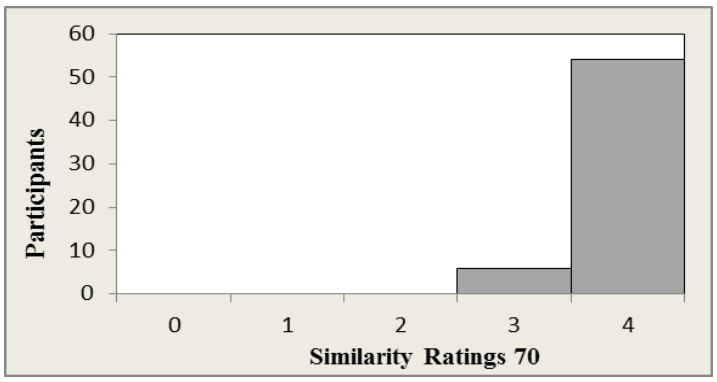


Figure 5.4 Histogram of similarity ratings for noun pair 70, SD= 0.28.

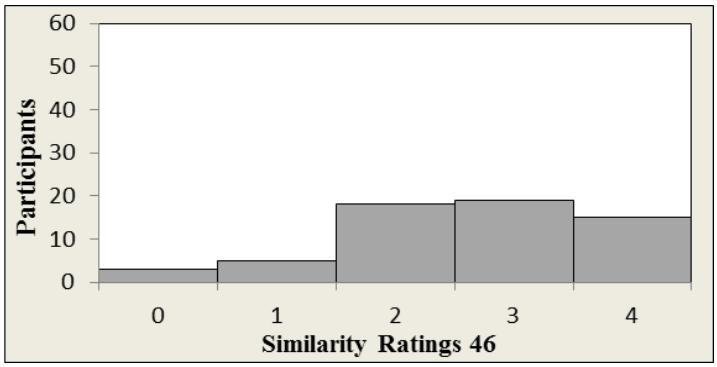


Figure 5.5 Histogram of similarity ratings for noun pair 46, SD= 1.07.

5.2.4.2 Comparison with the R&G Dataset

The R&G dataset was used as a general framework for the production of the ANSS-70 dataset. In this section, a comparison has been conducted between the two datasets to illustrate the differences between them.

1. ***Method of Selection of Materials***

R&G used a set of 48 nouns to make up a combination of 65 noun pairs spanning the range of semantic similarity values from maximum to minimum. This dataset was published without justification for the specific choices of 48 nouns and the method used to make up of the combination of 65 noun pairs. The R&G dataset is skewed towards low similarity word pairs (Miller and Charles, 1991).

This study used a set of 56 stimulus Arabic nouns that were carefully selected through the use of 27 Arabic categories created for 27 themes. Semantic similarity judgments are an issue of human perception. An experiment was conducted to make up a combination of 70 noun pairs spanning the similarity range based on human judgments to counter the bias towards low similarity in the R&G dataset.

2. ***Sampling the Population of Participants***

The sample of participants used in the R&G experiment for the collection of human ratings consisted of two groups of college undergraduates with a total of 51 participants. No information was provided as regards the composition of age or gender for each group and whether the sample of participants used in this experiment contained only native English speakers.

The sample of human population used in the ANSS-70 dataset experiments is more representative than the R&G experiment. The value of a sample of participants selected to carry out a specific experiment could be reduced as a representative sample if there is a high homogeneity of participants and they are distant from the general population. Consequently, the sample was selected as a general population (students and non-students) from different Arabic countries taking into account the gender, age, and academic background of the participants. The sample was selected to balance gender (males and females), student and non-student, academic background (science/engineering vs. arts/humanities) and age to avoid any possible bias.

3. *The Procedure of Collection of Human Ratings*

A card sorting technique was used for collecting human ratings in the R&G experiment. Each of the noun pairs was printed on a separate slip and the order of the 65 slips was randomized before presentation. The participants were asked to sort the slips into order of similarity of meaning and each noun pair was rated by assigning a value from 4.0- 0.0: "the greater the similarity of meaning the higher the number". These instructions concentrate on the relative similarities of noun pairs in the dataset. This may encourage expansion of the range of similarity ratings to fill the range 4.0 to 0.0, regardless of whether other noun pairs with higher or lower similarity exist external to the dataset (O'Shea, 2010).

A combination of card sorting with semantic anchors was used to collect human ratings in the ANSS-70 dataset experiment. Using the semantic anchors could offer better interval measurement and also lower noise than R&G method whereby more consistent human ratings (lower noise) was demonstrated by this combination notably as regards the unsupervised collection of ratings from the general population sample. Each noun pair in the Arabic noun dataset was printed on a separate card and the order of 70 cards was randomized before presentation. The participants were asked to sort the cards into four groups based on the similarity of meaning. The noun pairs in each group were rated using a point rating scale (the points described by the semantic anchors) which ran from 0 (low similarity) to 4 (high similarity).

5.2.5 Evaluation Procedure

5.2.5.1 Creation of Evaluation and Training Sub-Datasets

The evaluation process of the Arabic noun similarity (KalTa-A) measure required identifying its optimal parameter values. Therefore, the ANSS-70 dataset has been divided into two sets. The first known as the training dataset was employed to tune the KalTa-A measure parameters whilst the second denoted as evaluation dataset was used to assess its accuracy. Each dataset consisted of 35 noun pairs spanning the similarity of meaning range from maximum to minimum, which were selected as follows.

1. The original ANSS-70 dataset consisted of 24 low similarity, 24 medium similarity and 22 high similarity noun pairs. Therefore, each sub-dataset contained 12 low similarity, 12 medium similarity and 11 high similarity noun pairs.
2. For each similarity class within the same sub-dataset, the noun pairs were selected with similarity of meaning ranging from low to high. Figures 5.6 and 5.7 present the noun pairs in the evaluation dataset and training dataset respectively.

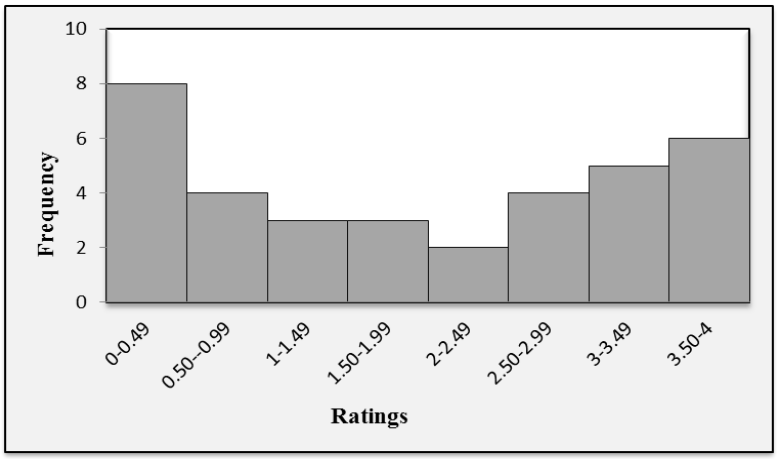


Figure 5.6 Distribution of similarity ratings in the evaluation dataset

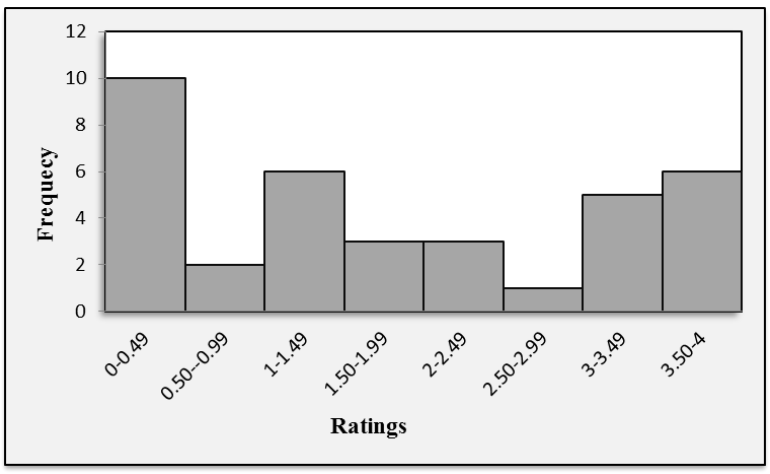


Figure 5.7 Distribution of similarity ratings in the training dataset

Only 30 noun pairs of each sub-datasets have been used in the training and evaluation experiments on account of the fact that some Arabic nouns have not been added to the current version of AWN such as موقد stove, ساحر wizard, مستشفى hospital… etc. In addition, some Arabic nouns do not have complete senses such as

the Arabic word ضحك laugh, which has just two senses in the current version of AWN while the sense (laugh as a facial expression) has not been added to the current version.

The noun pairs in the training and evaluation datasets are listed with human ratings in table 5.9 and table 5.10, respectively. The faint grey noun pairs have not been used in the training and evaluation experiments.

Table 5.9 Training Dataset Noun Pairs with Human Ratings

No.	Word Pairs		Human Ratings	أزواج الكلمـــات	
1	Cushion	Diamond	0.01	الماس	مسند
2	Gem	Pillow	0.02	مخدة	جوهرة
3	Cord	Midday	0.02	ظهيرة	حبل
4	Signature	String	0.02	خيط	توقيع
5	Boy	Endorsement	0.03	تصديق	صبي
6	Boy	Midday	0.04	ظهيرة	صبي
7	Smile	Pigeon	0.05	حمامة	ابتسامة/بسمة
8	Noon	Fasting	0.07	صيام	ظهر
9	Countryside	Laugh	0.08	ضحك	ريف
10	Glass	Fasting	0.10	صيام	كأس
11	Hospital	Grave	0.21	قبر	مستشفى
12	Gem	Young woman	0.22	شابة	جوهرة
13	Run	Shore	0.28	شاطئ	جري
14	Hill	Woodland	0.30	أحراش	تـل
15	Countryside	Vegetable	0.31	خضار	ريف
16	Master	Thinker	0.34	مفكر	سيد
17	Feast	Laugh	0.34	ضحك	عيد
18	Hen	Oven	0.36	فرن	دجاجة
19	Slave	Lad	0.44	فتى	عبــد
20	Journey	Bus	0.46	باص	رحلة
21	Girl	Odalisque	0.49	جارية	فتاة
22	Brother	Lad	0.54	فتى	أخ
23	Sage	Sheikh	0.57	شيخ	حكيم
24	Hen	Pigeon	0.65	حمامة	دجاجة
25	Brother	Sister	0.77	أخت	أخ
26	Sage	Thinker	0.83	مفكر	حكيم
27	Gem	Diamond	0.85	الماس	جوهرة
28	Journey	Travel	0.85	سفر	رحلة
29	Smile	Laugh	0.87	ضحك	ابتسامة/بسمة
30	Stove	Oven	0.89	فرن	موقد
31	Signature	Endorsement	0.90	تصديق	توقيع
32	Noon	Midday	0.93	ظهيرة	ظهر
33	Girl	Young Woman	0.94	شابة	فتاة
34	Coach	Bus	0.95	باص	حافلة
35	Hospital	Infirmary	0.98	مشفى	مستشفى

5.2.5.2 Tuning Parameters

The KalTa-A measure parameters (α and β) have been tuned using the training dataset in order to determine the optimal values within the interval [0, 1]. Given the initial value of each parameter, the training dataset noun pairs were run using the KalTa-A measure to produce machine similarity ratings in a range of 0 to 1. The correlation coefficient between the human ratings of training dataset and those obtained from the KalTa-A measure was computed. The values of the Arabic measure parameters were changed to obtain a set of correlation coefficients. The increasing step of α and β was 0.05. Then the parameters with the strongest correlation coefficient were considered as the optimal parameters. In this experiment, the strongest correlation coefficient was obtained at $\alpha = 0.12$ and $\beta = 0.21$.

Using the identified optimal parameters, the noun pairs on the evaluation dataset were run to produce the machine similarity ratings. The correlation coefficient was calculated again between the machine and human ratings for pairs of nouns on the evaluation dataset to assess the accuracy of the KalTa-A measure.

The KalTa-A measure calculated the similarity between two Arabic nouns using the AWN and SUMO ontologies as described in chapter 4. For the purpose of comparison, the tuning parameters procedure was performed again to identify the optimal parameter values for KalTa-A measure without SUMO (using the AWN only). The optimal values were $\alpha = 0.162$ and $\beta = 0.234$.

The machine similarity ratings were produced by running the evaluation dataset noun pairs on the KalTa-A measure without SUMO using the identified optimal parameters. Table 5.10 shows the human similarity ratings with the corresponding machine similarity ratings on the evaluation dataset. The first and last columns represent the noun pairs on the evaluation dataset in Arabic and English. The second column represents the human similarity ratings which were rescaled from 0 - 4 to 0 – 1 for the purpose of comparison. The third and fourth columns represent the corresponding machine similarity ratings produced by the KalTa-A measure without SUMO and the KalTa-A measure respectively. The faint grey noun pairs have not been used in the evaluation experiments.

Table 5.10 Evaluation Dataset Noun Pairs with Machine and Human Ratings

No.	Word Pairs		Human Ratings	KalTa-A without SUMO	KalTa-A Ratings	أزواج الكلمـــات	
1	Coast	Endorsement	0.01	0.0	0.12	تصديق	ساحـل
2	Noon	String	0.01	0.27	0.31	خيـط	ظهـر
3	Stove	Walk	0.02	-	-	مشـي	موقــد
4	Slave	Vegetable	0.04	0.06	0.08	خضـار	عبد
5	Smile	Village	0.05	0.0	0.10	قريــة	ابتسامة/بسمة
6	Wizard	Infirmary	0.06	-	-	مشفى	ساحـر
7	Hill	Pigeon	0.08	0.06	0.10	حمامة	تـل
8	Glass	Diamond	0.09	0.05	0.07	الماس	كأس
9	Cord	Mountain	0.13	0.17	0.20	جبـل	حبـل
10	Forest	Shore	0.21	0.17	0.20	شاطئ	غابــة
11	sepulcher	Sheikh	0.22	0.06	0.08	شيخ	ضريـح
12	Tool	Pillow	0.25	0.32	0.35	مخـدة	أداة
13	Coast	Mountain	0.27	0.45	0.48	جبـل	ساحل
14	Tool	Tumbler	0.33	0.54	0.60	قـدح	أداة
15	Journey	Shore	0.37	0.0	0.25	شاطئ	رحلة
16	Coach	Travel	0.40	0.0	0.54	سفر	حافلة
17	Food	Oven	0.44	-	-	فـرن	طعـام
18	Feast	Fasting	0.49	0.17	0.20	صيـام	عيـد
19	Coach	Means	0.52	0.38	0.43	وسيلـة	حافلة
20	Girl	Sister	0.60	0.37	0.44	اخـت	فتــاة
21	Hill	Mountain	0.65	-	-	جبــل	تــل
22	Master	Sheikh	0.67	0.67	0.71	شيخ	سيـد
23	Food	Vegetable	0.69	0.53	0.54	خضـار	طعـام
24	Slave	Odalisque	0.71	0.93	0.90	جاريـة	عبـد
25	Run	Walk	0.75	0.60	0.62	مشي	جـري
26	Cord	String	0.77	0.70	0.70	خيـط	حبـل
27	Forest	Woodland	0.79	0.82	0.78	أحراش	غابـة
28	Cushion	Pillow	0.85	0.82	0.78	مخدة	مسند
29	Countryside	Village	0.85	0.82	0.78	قرية	ريف
30	Coast	Shore	0.89	0.89	0.85	شاطئ	ساحل
31	Tool	Means	0.92	0.93	0.90	وسيلة	أداة
32	Boy	Lad	0.93	0.95	0.93	فتى	صبي
33	Sepulcher	Grave	0.94	0.82	0.78	قبـر	ضريح
34	Wizard	Magician	0.94	-	-	مشعوذ	ساحـر
35	Glass	Tumbler	0.95	0.89	0.85	قـدح	كـأس

5.2.6 Findings and Discussion

The possible indicative value and bounds of a performance expected from the KalTa-A measure were calculated as the average, worst and best performances of human participants on the evaluation dataset as shown in table 5.11. This was undertaken

using the leave-one-out resampling technique in order to calculate the correlation coefficient of each of 60 participants with the rest of the group. The correlation coefficient was calculated for each of the 60 participants between the participant's ratings and the average ratings of the rest of the group. The consistency of the KalTa-A measure with human perception was identified by computing the correlation coefficient between the average rating of human participants and the machine ratings as shown in table 5.11.

Table 5.11 The Performance of KalTa-A measure on the Evaluation dataset.

On Evaluation Data Set	**Correlation *r***
KalTa-A measure	0.91
KalTa-A measure without SUMO	0.894
Average of the correlation of all participants	0.893
Best participants	0.970
Worst participants	0.716

The KalTa-A measure without SUMO obtained a good value of the Pearson correlation coefficient (r = 0.894) with the human judgments as shown in figure 5.8. The KalTa-A measure without SUMO is performing well at (r = 0.894) with the average value of the correlations of human participants (r = 0.893). Furthermore, the performance of the Arabic measure is substantially better than the worst human (lower bound) performance at (r =0.716).

As mentioned in chapter 4, the KalTa-A without SUMO measure ratings were hampered by the structure of the AWN noun hierarchy which may produce a bias towards a particular distance computation such as the noun pairs 15 and 16 in table 5.10. These pairs were rated medium by participants whilst very low similarity values obtained by the KalTa-A measure without SUMO. An explanation is provided by consideration of the noun hierarchy in AWN. The nouns of the pair 15 رحلة شاطئ (*Journey* and *Shore*) are classified under separate substructures which show no connection between them in the AWN noun hierarchy leading to the obtainment of a very low similarity value by the KalTa-A measure. The noun pair 16 حافلة سفر (*Coach* and *Travel*) obtained a machine rating lower than the human similarity rating for similar reasons

The performance of the KalTa-A measure improved using SUMO which achieved a correlation (r = 0.91) better than the correlation obtained by the KalTa-A measure without SUMO at (r = 0.894) as shown in table 5.11. The machine similarity ratings of the noun pairs 15 and 16 were improved using the SUMO. Whereby, medium similarity values were obtained for the noun pairs 15 and 16 which were very close to the human assessment as shown in table 5.10. Figure 5.9 shows the correlation between the KalTa-A measure and human ratings.

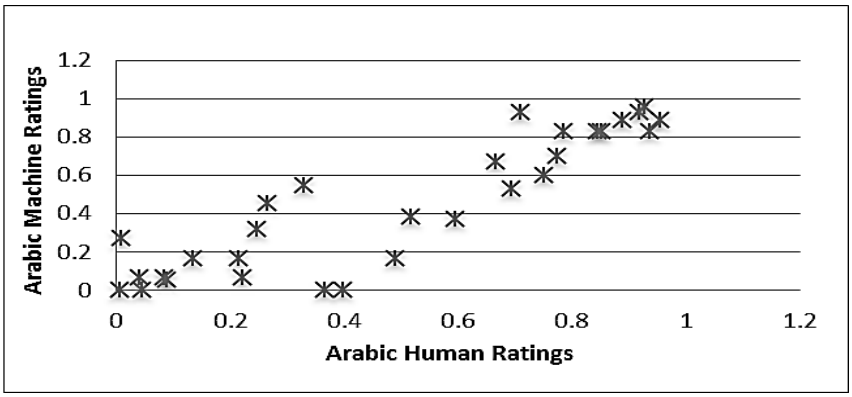


Figure 5.8 The Correlation between the Ratings of Human and the KalTa-A measure without SUMO

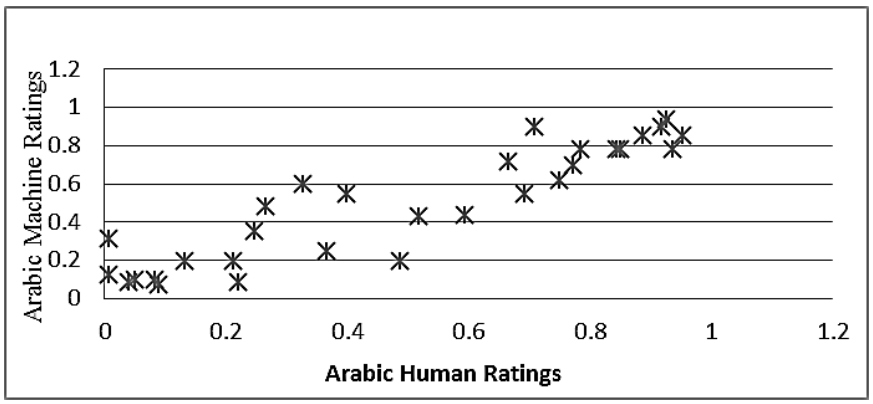


Figure 5.9 The Correlation between the Human Ratings and the KalTa-A measure

5.3 Creating an Arabic Verb Benchmark Dataset (AVSS-70)

This section describes the production of a new Arabic verb benchmark dataset, namely that of the AVSS-70. It is the first of its kind for Arabic which was particularly developed to assess the accuracy of Arabic verb semantic algorithms.

The stages of the AVSS-70 dataset process design were adopted from the ANSS-70 dataset creation procedure which included selection of the stimulus verbs, production of representative pairs of verbs and collection of the human similarity ratings.

The AVSS-70 dataset adapted elements from work on English text semantic similarity to select stimulus verbs which represented the verbs in the Arabic language. Once again there was novel work in generating a set of verb pairs which provides the best representation (for its size) of the huge range of verb pairs that can be generated from the verbs in the Arabic language.

5.3.1 Selecting the Stimulus Verbs

Representation of the verbs in the Arabic language was achieved by carefully selecting 25 stimulus verbs by means of adaption of a sampling frame technique that used by (O'Shea et al., 2013) to create a short text dataset for English. The sampling frame is a method of representing a large population with a small carefully-chosen sample randomly selected with constraints. Selecting the stimulus verbs consisted of two steps including:

1. Decomposing the Arabic verbs into a hierarchy of classes
2. Populating the slots in the frame with verbs using random selection where choice is possible.

5.3.1.1 Decomposing the Arabic verbs into a hierarchy of classes

In this research, the Arabic verbs were decomposed into a tree structure based on special syntactical and semantic features. Each of the tree levels is described in this section.

Most theoretical work on Arabic verb classes is based on the root and template based method (Mousser, 2010). It was decided not to apply this method as it was used by the KalTa-F measure to calculate the similarity between two Arabic verbs and this would avoid biasing the AVSS-70 dataset in favour of the Arabic verb algorithm. An alternative method was needed at this stage and a set of more sophisticated grammatical techniques developed for NLP such as Case Grammar (CG) and Arabic

VerbNet (AVN) were applied instead. AVN was inspired by Levin classes (Levin, 1993).

Case Grammar

(Al-Qahtani, 2005) presented an extensive classification of Arabic verbs based on Case Grammar (CG) as described by (Fillmore, 1968). The classification was based on Cook's Matrix Model (Cook, 1979) and its extension.

CG classified the Arabic verbs into three classes comprising state, process and action which are useful in a high-level decomposition. The top-level breakdown of the Arabic verbs is shown in figure 5.10.

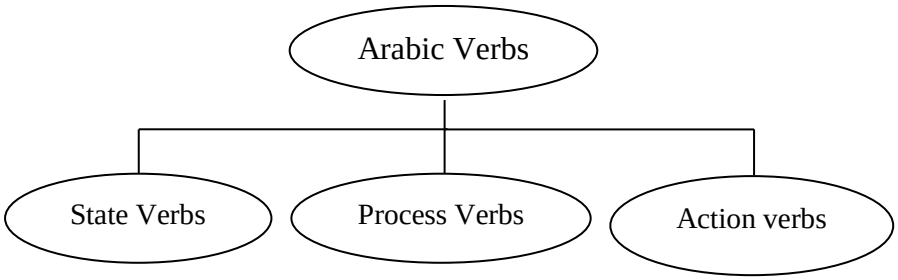


Figure 5.10 Top level Arabic verbs decomposition

Each verb class was decomposed further into basic, experiential, benefactive, and locative verbs which offered a good intermediate level, as shown in figure 5.11.

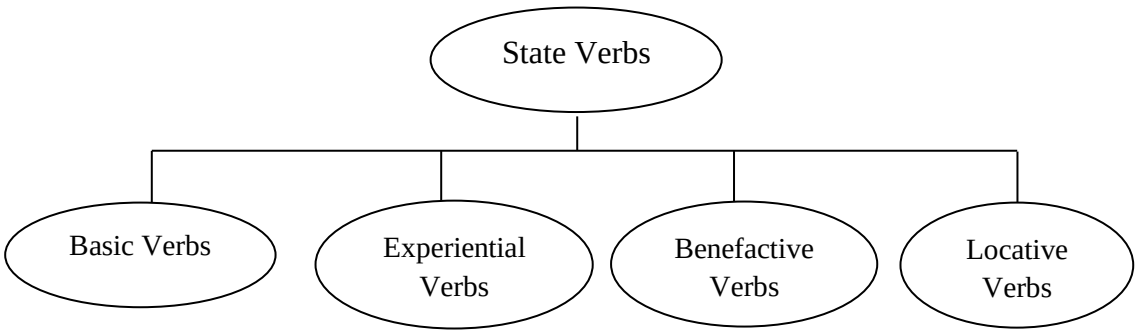


Figure 5.11 The decomposition of the state verbs at intermediate level.

(Al-Qahtani, 2005) described 5 cases used by Cook which represent all propositional cases required by the semantic valence of the verb. These comprised the Object case (O) which is always found with every verb, Agent (A) which is the case needed by

an active verb, Experiencer (E) which is the case needed by an experiential verb, the Benefactive (B) case which is required by a benefactive verb and the Locative (L) case required by a locative verb.

Each of the 12 classes at intermediate level contains verbs occurring with one, two or three cases of the 5 Cook's cases based on CG. Table 5.12 illustrates this. For example, there is only one case (object) with the verb class number 1(State: Basic) such as the verb صدر "be issued" in the sentence صدر الجزء الاول "*the first volume was issued*". Two cases, however, (E, O_S) occur with the verb class number 2 (State: Experiential): for instance, the verb رأى "see" in the sentence رأى زيد الحادثة " *Zayd saw the accident*" where Zayd is the experiencer (E) and accident is the object (Al-Qahtani, 2005).

Table 5.12 The 12 Arabic verb classification based on case grammar

Class	Verb Types	Case Frames
1	State Basic	Os
2	State Experiential	E, Os
3	State Benefactive	B, Os
4	State Locative	L, Os
5	Process Basic	O
6	Process Experiential	E, O
7	Process Benefactive	B, O
8	Process Locative	O, L
9	Action Basic	A, O
10	Action Experiential	A,E,O
11	Action Benefactive	A,B,O
12	Action Locative	A,O, L

(Al-Qahtani, 2005) classified Arabic verbs based on Cook's Matrix Model (12 classes) and its extension. In this dataset, Cook's model extension was considered for further decomposition of Arabic verbs. Each of the 12 classes was extended into two case frames to include:

- Frame with a time (T) case: for example, "*they spent 3 days in Paris*", the verb *spend* requires an essential time in its semantic valence.
- Double object (O, O) case frame: for example the verb *appoint* in the sentence *he appointed him in his company*, takes one object while the verb *name* takes two objects in addition to the agent in this example, *He names his child Ali*.

- Frames different in subject choice: for example, *Wrestling excites Zayd*, the object (O) of the verb *excite* in this sentence appears first and the subject (Experiencer) follows (O, E) whilst in the sentence *Zayd is excited by wrestling*, the subject is first and the object is second (E, O).

Table 5.13 shows the final results of Arabic verb decomposition based on CG which consist of 24 case frames. These case frames were employed in the next stage decomposition which offered the capacity for further fine-grained decomposition.

Table 5.13 Arabic Verb Classification based on Case Grammar

Class	Verb Types	Case Frames
01	State Basic	Os
02	State Basic	Os, Os
03	State Experiential	E, Os
04	State Experiential	Os, E
05	State Benefactive	B, Os
06	State Benefactive	Os, B
07	State Locative	L, Os
08	State Locative	Os, L
09	Process Basic	O
10	Process Basic	O, O
11	Process Experiential	E, O
12	Process Experiential	O, E
13	Process Benefactive	B, O
14	Process Benefactive	O, B
15	Process Locative	O, L
16	Process Locative	L, O
17	Action Basic	A, O
18	Action Basic	A,O,O
19	Action Experiential	A,E,O
20	Action Experiential	A,O,E
21	Action Benefactive	A,B,O
22	Action Benefactive	A,O,B
23	Action Locative	A O,L
24	Action Locative	A,L,O

Arabic Verbnet (AVN)

(Mousser, 2010) presented a large coverage verb lexicon for the Arabic language which exploited Levin's verb-classes (Levin, 1993) with the development procedure described by (Schuler, 2005). The largest English verb classification is that of Levin's classes which classified English verbs into groups based on syntactic properties and the verb's meanings which are related but not necessarily synonymous

(Kipper et al., 2000). The hierarchal Arabic lexicon has been built based on the notion that Verb Classes idea can be transferred into Arabic with some adaptations. Members of each class have been translated into Arabic by means of applying Levin's class inventory in Arabic. This process showed that many Levin classes do not exist in Arabic and also that the event structures of some Arabic verbs have not been described by Levin's class inventory. Consequently, some Levin classes have been integrated into other classes, some Levin classes have been divided into two classes and some new classes and sub-classes have been created. This work produced good verb classes which were used in the final stage decomposition in this research.

Combining CG and AVN classes for decomposition offered a good intermediate structure and fine-grained classes which were easy to understand and use. In the final level of Arabic verb decomposition, each of the 24 CG verb classes (table 5.13) at the intermediate level was combined with a different class from the AVN verb classes at different levels. Figure 5.12 shows a portion of the Arabic verb tree structure where the case frame (State: Benefactive: O_S B) was combined with the top level AVN verb class (**IiDotar~a-1**) whilst the case frame (State: Benefactive: B O_S) was combined with the third level AVN verb class (**Ii$otaray-1.1.1**).

Consequently, the number of slots (stimulus verbs) selected to create the dataset verb pairs were 25. It was decided to allocate 24 slots to the 24 CG verb classes presented in table 5.13. There is no case frame that represents the frame with a time case (T) in 24 CG verb classes. Therefore, it was decided to allocate slot number 25 as the verb with the time (T) case frame.

Each of 25 slots would be also allocated to a different AVN verb class to promote semantic dispersion, where 20 slots would be allocated to the top level of the AVN verb classes and 5 slots would be allocated to the lower level AVN verb classes (second and third).

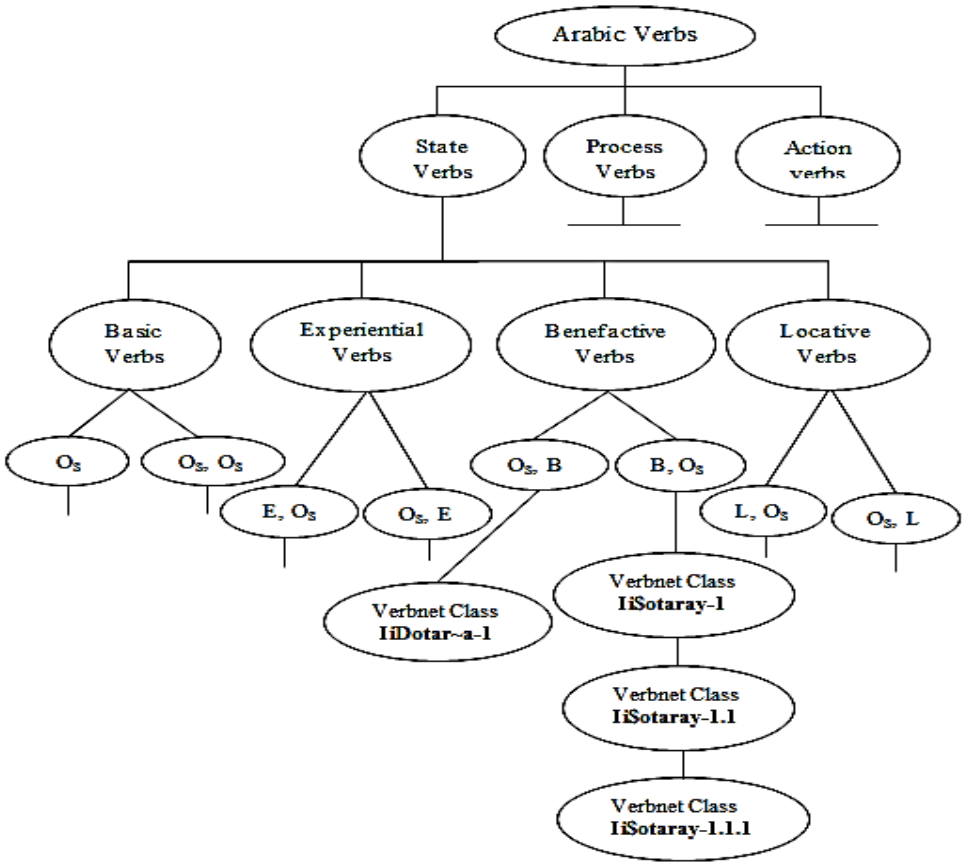


Figure 5.12 A Portion of Arabic Verbs Tree Structure.

5.3.1.2 Population of the Slots in the Frame with Arabic Verbs

Another important issue in language representation is that of word frequency i.e. high frequency verbs should have a higher probability of appearing in the sample frame. For valid verb representation, verbs which are selected to fill the slots in the frame are in proportion to their frequency of appearance.

Consequently, the decision was made to use rule 80/20 used by (O'Shea et al., 2013) whereby 80% of the slots in the frame will be filled by random selection process with words from the high frequency word list whilst 20% from the list of low frequency words. This approach concentrated on the core vocabulary used in teaching language. (Valcourt and Wells, 1999) reported that 80% of undergraduate English textbook words are a high frequency set.

The set of stimulus verbs will be presented to the general population which requires selecting ordinary Arabic verbs (high frequency). However, for valid verb

representation, 80% of 25 stimulus verbs (20 verbs) will be selected from the high frequency Arabic word list and 20% (5 verbs) from the low frequency word list.

A number of studies have been carried out over the years by the Arab and non- Arab researchers (Al-Azawi and Buradah, 1976, Lutfi, 1948) in order to compile lists of Arabic high frequency words to be used in teaching language. Quitregard (1994) listed the vocabulary of the most common 2000 Arabic words which could be utilised to teach language. Unlike other studies, the texts of this study were compiled from various Arabic sources and from different Arabic countries. The texts were derived from various types of publications (such as drama, essays, fiction, geographical, historical and scientific works), newspapers from 14 countries, magazines from 9 countries, films and television programmes from 7 countries, radio programmes from 12 countries, literary histories and children's books.

The most up to date work is a list of 37,000 Arabic words ranked by frequency known as Arabic Word Count (AWC) (Attia et al., 2011) and this was partitioned into 2000 high frequency word list and the remainder as low frequency word list on the basis of that the most frequent 2000 words is the core of the language. Thus, 80% of the stimulus verbs are selected from the 2000 high frequency word list and 20% from the low frequency word list.

The next step in the process of filling the slots included creating a list of high frequency verbs by means of selecting all verbs from the list of (2000) high frequency words, whereas the low frequency verb list contained all verbs from the list of low frequency words (the remainder of the AWC list). The verbs in each list were randomised to avoid the occurrence of bias. Each list was searched to select verbs to fill the slots based on their criteria as specified through the process of verb decomposition. The following steps illustrate the process of filling the slots:

1. For the high frequency verb list
2. Select the first verb
3. When the verb meets the criteria for any high frequency slot, allocate the verb to the slot
4. Otherwise discard the verb
5. Select the next verb

6. Repeat steps 2 to 6, until all the high frequency slots are filled

Once 80% of the slots have been filled, switch to the low frequency list and repeat the procedure from 2 to 6 until all the low frequency slots are filled. Table 5.14 illustrates the results of the process of populating the frame, where LF means the verb selected from the low frequency verb list.

Table 5.14 Populated verb sampling frame

No.	Verb classes		Selected verb	الفعل	VerbNet class
01	State verb basic	Os	Be capable	تمكن	**Najaha-1.2**
02	State verb basic	Os, Os	Include	تضمن	Oaqal~a-1
03	State verb experiential	E, Os	Believe	اعتقد	Saw~ara-1
04	State verb experiential	Os, E	Appear	بدا	Zahara-1
05	State verb benefactive	Os, B	Be forced	اضطر	IiDotar~a-1
06	State verb benefactive	B, Os	Get	نـال	**Ii$otaray-1.1.1**
07	State verb Locative	Os, L	Contact	اتصل	LaAqy-1
08	State verb locative	L, Os	Overcrowd **LF**	اكتظ	Iimotalaoa-1
09	Process verb Basic	O	Increase	ازداد	TadaEafa-1
10	Process verb Basic	O, O	Become	صـار	Saara-1
11	Process verb experiential	E, O	Hope	تمنى	OaraAda-1
12	Process verb experiential	O, E	Happen	حدث	Hasala-1
13	Process verb benefactive	B, O	Find **LF**	لقي	Wujada-1
14	Process verb benefactive	O, B	Enrich **LF**	اثرى	Other-cos-1
15	Process verb Locative	O, L	Go	ذهب	Haraba-1
16	Process verb Locative	L, O	Leak **LF**	نضح	**Nazafa-1.1.1**
17	Process verb Time	O, T	Continue	استمر	Iisotamar~a-1
18	Action verb basic	A, O	Try	حاول	HaAwala-1
19	Action verb basic	A,O,O	Appoint	عين	Eay~ana-1
20	Action verb experiential	A,E,O	Announce	اعلن	Eab~ara-1
21	Action verb experiential	A,O,E	Allow	سمح	Samaha-1
22	Action verb benefactive	A,B,O	Accept	قبل	Taqab~ala-1
23	Action verb benefactive	A,O,B	Give	اعطى	**OaEotay-1.1**
24	Action verb locative	A O,L	Arrive	وصل	Haraba-1
25	Action verb locative	A,L,O	Fill **LF**	ملأ	**Gamara-1.1**

Some problems arose in the process of filling the slots. For example, the verb "include" تضمن was selected from the list of high frequency verb because it met the criteria of the slot number 02 in table 5.14 (State verb: Basic: Double object). The

slot number 02 was also allocated to a high level AVN class however the verb "include" تضمن has not been added to AVN yet. To solve this problem, this verb was allocated to the high level AVN class Oaqal~a-1 which contains verbs sharing a meaning component such as comprise شمل, contain احتوى, etc.

5.3.2 Constructing the Set of Arabic Verb Pairs

A new method was used for generating a representative sample of 70 verb pairs based on human judgements. The sample size was chosen based on experience gained from the previous experiment in constructing the ANSS-70 dataset which indicated that the sample of 70 noun pairs was sufficient for evaluating the KalTa-A algorithm.

The ANSS-70 dataset creation process has shown that high similarity pairs are harder to specify than one might anticipate. It has also shown that is hard to predict where proposed medium similarity pairs might lie on the scale. Where 24 candidate noun pairs were selected for the high similarity range but only 22 noun pairs were rated high by the 60 participants and the rest (2 pairs) were rated as medium. Therefore there are slightly more high similarity pairs than medium and low similarity on the expectation that some high and medium similarity pairs will be in the band below. The steps of the creation of a representative sample of Arabic verb pairs are described in the following sections.

5.3.2.1 High Similarity Verb Pairs

The set of high similarity verb pairs should contain pairs between strongly related in meaning and identical in meaning. It was decided to make use of AVN classes (Mousser, 2010) for producing this set as the verbs in the AVN were classified into classes based on shared meaning and behaviour. Verbs such as leave غادر, desert هجر, quit ترك, depart برح etc. share a meaning component and were grouped into a verb class denoted as the GAADARA-1 class. For each verb in the list of stimulus verbs, another verb was selected which was paired with it in the same AVN class. Again selection was adjusted to achieve an overall 80% high frequency, 20% low frequency. For example, the stimulus verb *include* تضمن was grouped with some

verbs in the verb class Oaqal~a-1 which included verbs such as contain احتوى, comprise شمل, accommodate استوعب etc. The verb *comprise* شمل was selected to pair with the verb *include* تضمن based on its frequency of appearance in the high frequency list. The set of high similarity verb pairs is presented in table 5.15. English translations are approximation and also Arabic word may have different sets of polysemous senses to corresponding English words (e.g. as in R&G, glass and tumbler each have sets of polysemous senses).

Table 5.15 The High Similarity Verb Pairs

No.	High similarity verb pairs		ازواج التشابه العالي		AVN verb class
01	Be capable	Be able	استمكن	تمكن	Najaha-1.2
02	Include	Comprise	شمل	تضمن	Oaqal~a-1
03	Believe	Consider	اعتبر	اعتقد	Saw~ara-1
04	Appear	Appear	ظهر	بدا	Zahara-1
05	Be forced	Be obligatory	وجب	اضطر	IiDotar~a-1
06	Get	Obtain	حصل	نـال	Ii$otaray-1.1.1
07	Contact	Meet	التقى	اتصل	LaAqy-1
08	Overcrowd	Crowed	ازدحم	اكتظ	Iimotalaoa-1
09	Increase	Rise	ارتفع	ازداد	TadaEafa-1
10	To be	Become	اصبح	صـار	Saara-1
11	Hope	Wish	رغب	تمنى	OaraAda-1
12	Happen	Take place	جرى	حدث	Hasala-1
13	Find	Find	وجد	لقي	Wujida-1
14	Enrich	Richen	اغنى	اثرى	Other-cos-1
15	Go	Depart	غادر	ذهب	Haraba-1
16	Leak	Seep	تسرب	نضح	Nazafa-1.1.1
17	Continue	Go on	واصل	استمر	Iisotamar~a-1
18	Try	Endeavour	سعى	حاول	HaAwala-1
19	Appoint	Employ	وظف	عين	Eay~ana-1
20	Announce	Declare	صرح	اعلن	Eab~ara-1
21	Allow	Permit	اجاز	سمح	Samaha-1
22	Accept	Approve	وافق	قبل	Taqab~ala-1
23	Give	Grant	منح	اعطى	OaEotay-1.1
24	Arrive	Come	جاء	وصل	Haraba-1
25	Fill	Load	عبأ	ملأ	Gamara-1.1

5.3.2.2 Medium Similarity Verb Pairs

Unlike high and low similarity verb pairs, it was relatively difficult to generate a set of medium similarity verb pairs which consisted of pairs between vaguely similar and very much alike in meaning. A new approach was used to generate a set of medium verb pairs based on human judgement which required the following steps to be completed:

1. Selection of an original verb and the use of participants to create a list of 8 synonyms.

2. Selection of one verb from the list of 8 synonyms as a stimulus verb and the use of participants to create a new list of 8 synonyms.

3. Convening a committee of 4 judges to select appropriate pairing with the original verb as medium similarity from the two lists of synonyms created in step1 and step2.

5.3.2.2.1 Creation of the List of Original Verbs (LOV)

The primary aim of the medium similarity verb pairs' generation method was to create lists of synonyms by participants and utilise them to select a set of medium verb pairs. This required creating a List of Original Verbs (LOV) which was presented to the participants who were requested to create a list of synonyms for each original verb. LOV was created through the random selection of 23 verbs from 50 verbs that were used to make up the set of 25 high similarity verb pairs. Each of 50 verbs was printed on a slip of paper. The 50 slips were mixed and 23 slips were selected randomly. Table 5.16 illustrates the list of 23 verbs which was employed in the next step of medium similarity verb pairs' generation.

Table 5.16 The List of Original Verbs (LOV)

	Selected verbs	الافعال
1	Be capable	تمكن
2	Include	تضمن
3	Consider	أعتبر
4	Appear	ظهر
5	Be obligatory	وجب
6	Obtain	حصل
7	Contact	اتصل
8	Crowd	أزدحم
9	rise	إرتفع
10	Hope	تمنى
11	Happen	حدث
12	Find	وجد
13	Enrich	أغنى
14	Depart	غادر
15	Leak	تسرب
16	Continue	استمر
17	Try	حاول
18	Appoint	عين
19	Declare	صرح
20	Approve	وافق
21	Give	أعطى
22	Arrive	وصل
23	Fill	ملأ

5.3.2.2.2 Experiment 1: Creation of the Lists of Synonyms

This step consisted of conducting an experiment to create the lists of 8 synonyms to each verb in LOV based on human judgment.

Participants

A sample of 4 native Arabic speakers from different academic backgrounds took part in this experiment and consisted of: Arabic linguistic (Iraq), Science/Engineering (Saudi Arabia), Secondary school (Iraq) and one selected randomly from Science / Engineering (Egypt)for this experiment.

Materials

The participants were supplied with 3 sheets including instructions, recording and personal information sheets. Appendix 3 contains examples of experimental materials including the appendix 3.1 instruction sheet and the appendix 3.2 recording sheet.

The recording sheet contained a table with three columns. The first column contained the list of original verbs created in step 1whilst the second and third columns were used by the participants to write down two verbs for each of the original verbs. The basic instructions informed the participants that they would be requested to produce two verbs for each original verb on the recording sheet. The final sheet contained minimal details about the participant including name, age, degree title and a confirmation that the participant is an Arabic native speaker.

Procedure

Participants were asked to write down two verbs which could be used in the original verb's place in a sentence, i.e. means the same or very close in meaning. Participants were requested to write down two verbs as it was expected that they would write the first thing that came into their heads as their initial response. For example, for the original verb ***include*** تضمن, the first verb written by all participants as a first response was ***contain*** احتوى, however they wrote different verbs as a second response.

Some notes were included in the instruction sheet asking the participants to write the verbs in clear handwriting and to avoid writing the original verb or writing the same verb twice as an answer. The participants were also asked to write two verbs for each of the 23 original verbs as all uncompleted questionnaires must be ignored.

Experimental Results

The result of this experiment was 23 lists of 8 synonyms. These lists were employed to produce a new list of 23 verbs as stimulus verbs for use in the next experiment to

generate new lists of 8 synonyms. The new list of stimulus verbs was created as follows:

- Duplicated verbs written by more than one participant were removed from each list of synonyms produced in this experiment.

- It was decided to remove verbs that make up a high similarity with the original verb in order to maximise the chances of getting the lower and of the medium similarity bound. This was undertaken by extracting the senses of the original verb and the senses of each verb in the list of synonyms from a well-known Arabic dictionary (Baalbaki, 2005). Each verb in the list of synonyms shared one sense or more with the original verb was removed. For example, table 5.17 shows the list of synonyms for the original verb *be capable* تمكن. The verbs written by participants 3 and 4 were removed as they were duplicated verbs. The senses of the original and the rest of the verbs were extracted from the dictionary. The verbs *can* and *be able* were removed because they shared senses with the original verb.

Table 5.17 The list of synonyms produced by participants for the original verb ***be capable***

Original verb	**Participant 1**		**Participant 2**		**Participant 3**		**Participant 4**	
Be capable	Can استطاع	Overcome تغلب	Be able قـدر	Overpower ظفـر	Can استطاع	Be able قــدر	Can استطاع	Be able قـدر

- One verb was randomly selected from the rest of verbs in each list of synonyms. Consequently, a list of 23 verbs was selected from 23 lists of synonyms to be used as stimulus verbs for the next experiment. Table 5.18 shows the 23 verbs selected in this experiment.

Table 5.18 The New List of 23 Verbs Produced in Experiment 1.

	Original verbs	Selected verbs	الافعال المختارة
1	Be capable	**Overcome**	تغلب
2	Include	**Accommodate**	استوعب
3	Consider	**Think / cogitate**	تدبر
4	Appear	**Become evident**	وضح
5	Be obligatory	**Require**	تطلب
6	Obtain	**Seize**	استولى
7	Contact	**Convene / meet**	اجتمع
8	Crowd	**Narrow**	ضاق
9	Rise	**Elevate / Progress**	ارتقى
10	Hope	**Want**	اراد
11	Happen	**Occur**	صار
12	Find	**Win**	ظفر
13	Enrich	**Be content with**	اكتفى
14	Depart	**Desert**	فارق
15	Leak	**Flow**	سال
16	Continue	**Persevere**	واضب
17	Try	**Exert**	بذل
18	Appoint	**Record / Register**	سجل
19	Declare	**Reveal**	كشف
20	Approve	**Admit**	تقبل
21	Give	**Spend**	انفق
22	Arrive	**Catch**	لحق
23	Fill	**Overflow**	طفح

5.3.2.2.3 Experiment 2: Creation of New Lists of Synonyms

The aim of this experiment was to create new lists of synonyms using the new list of 23 stimulus verbs produced in experiment 1. A new sample of 4 participants was used in this experiment which included: Arabic linguistic (Iraq), Science/Engineering (Libya), Secondary school (Iraq) and one selected randomly from Art/Humanities (Saudi Arabia) for this experiment.

The participants were also supplied with 3 sheets as in appendix 3 but the recording sheet contained the new list of 23 stimulus verbs produced in experiment 1. The same procedure used in experiment 1 was followed to create lists of synonyms and the participants were asked to write two verbs which could be used in the original

verb's place in a sentence. The results of this experiment were 23 new lists of 8 synonyms.

5.3.2.2.4 Selection of a Set of Medium Similarity Verb Pairs

A committee of 4 judges was convened to select a set of medium verb pairs. The judges background were that of Arabic linguistics (Syria), Islamic studies (Iraq), religious teaching (Bahrain) and computer science / Arabic natural language processing (Iraq).

Each member of the committee was provided with printed materials which were created using the list of original verbs LOV (table 5.16). For each of the original verb in LOV, the lists of synonyms collected in experiment 1 and 2 were combined together and were allocated to the original verb which had been written for it. The judges selected the final set of medium similarity by undertaking two steps as follows:

1. For each of the original verbs in LOV, the judges removed the verbs from its list of synonyms which had a high similarity when paired with the original verb.

2. One verb was selected from the rest of verbs in the list of synonyms which had a medium similarity when paired with the original verb (medium verb pairs are between vaguely similar and very much alike in meaning).

The final set of 23 medium similarity verb pairs is presented in table 5.19.

5.3.2.3 Low Similarity Verb Pairs

The set of 22 low similarity verb pairs were selected randomly. For each verb that was used to produce the sets of high and medium similarity verb pairs, the frequency of appearance of this verb in these sets was calculated. The verbs which occurred more than twice were removed to avoid a biased set of verbs from being used. The

remaining Arabic verbs were employed to randomly generate a set of low similarity verb pairs. High and medium similarity pairs already found were removed. The remaining pairs were selected at random as they were good candidates for low similarity. Table 5.20 illustrates the set of 22 low similarity verb pairs.

Table 5.19 The Set of Medium Similarity Verb Pairs

No	Medium Similarity Verb Pairs		ازواج التشابه المتوسط	
1	Be capable	Be superior	تفوق	تمكن
2	Include	Exist	وُجِد	تضمن
3	Consider	Ponder	تأمل	أعتبر
4	Appear	Find	وجد	ظهر
5	Be obligatory	Require	تطلب	وجب
6	Obtain	Realize	ادرك	حصل
7	Contact	Run across	قابل	اتصل
8	Crowd	Restrict	حصر	أزدحم
9	Rise	Richen	اغنى	إرتفع
10	Hope	Request	طلب	تمنى
11	Find	Take	اخذ	وجد
12	Enrich	Be strong	قوي	أغنى
13	Depart	Be far	ابتعد	غادر
14	Leak	Waste	هدر	تسرب
15	Continue	Work	اشتغل	استمر
16	Try	Want	اراد	حاول
17	Appoint	Specify	ثبت	عين
18	Declare	Explain	اوضح	صرح
19	Approve	Understand	تفهم	وافق
20	Give	Buy	اشترى	أعطى
21	Arrive	Catch up with	ادرك	وصل
22	Fill	Abound	كثر	ملأ
23	Happen	Find	وجد	حدث

Table 5.20 The Set of Low Similarity Verb Pairs

No	Low Similarity Verb Pairs		ازواج الحد الادنى للتشابه	
1	Be superior	Depart	غادر	تفوق
2	To be	Come	جاء	صـار
3	Waste	Explain	اوضح	هـدر
4	Include	Run across	قابل	تضمن
5	Become	Contact	اتصل	اصبح
6	Continue	Buy	اشترى	استمر
7	Leak	Be obligatory	وجب	نضح
8	Become	Be far	ابتعد	اصبـح
9	Be capable	Comprise	شمل	تمكن
10	Find	Permit	اجاز	لقي
11	Get	Seep	تسرب	نـال
12	Appear	Grant	منـح	بـدا
13	Overcrowd	Wish	رغـب	اكتـظ
14	Rise	Understand	تفهم	ارتفع
15	Fill	Declare	صـرح	مــلأ
16	Ponder	Load	عبـأ	تــأمل
17	Be forced	Enrich	اغنى	أضطر
18	Go	Believe	اعتقد	ذهـب
19	Try	Be far	ابتعد	حــاول
20	Enrich	Meet	التـقى	اثـرى
21	Require	Rise	ارتفع	تطـلب
22	Restrict	Appoint	عين	حصر

5.3.3 Collection of the Human Ratings Experiment

This experiment was conducted to collect human ratings for 70 pairs of verbs produced in section (5.3.2) using the card sorting technique with semantic anchors which was identified in the creation of the ANSS-70 dataset as a most suitable.

5.3.3.1 Participants

This experiment used a new sample of 60 participants. This sample was chosen based on experience with previous experiment of ANSS-70 dataset which indicated that the sample of 60 participants was sufficient for good quality ratings. The sample was selected as a general population with an equal balance between students and non-students.

- All were Arabic Native speakers from 10 Arabic countries including Iraq (15 participants), Saudi Arabia (15), Libya (8), Syria (6), Egypt (4), Palestine (4), Jordan (3), Morocco (2), Sudan (2), and Algeria (1).

- The participants' academic backgrounds consisted of 35 Science/Engineering vs. 21 Art/Humanities with 4 were secondary school. In case of educational level, the balance was obtained and the overall breakdown qualifications were illustrated in table 5.21.

Table 5.21 The Participants' Educational Qualification

Student	**Non-student (highest qualification)**
8 undergraduate	12 Bachelors
10 Masters	6 Masters
12 PhD	8 PhD
None	4 Secondary School

- In case of age, the average was 36 years with the standard deviation (SD) 8.3 years. Table 5.22 shows the age distributions of a selected sample.

Tale 5.22 Age distributions for Arabic population sample.

Age range	**Participants**	
20-29	14	Student 9
		Non-student 5
30-39	28	Student 16
		Non-student 12
40-49	14	Student 5
		Non-student 9
50-59	3	Student 0
		Non-student 3
60-69	1	Student 0
		Non-student 1

- The overall balance between female and male was achieved with 31 female and 29 male. Good gender balance was achieved for non-students with 16 female and 14 male while an equal gender balance was obtained from students (15 male and 15 female).

5.3.3.2 Materials

Each of 70 verb pairs was printed on a separate card to the same specification as the experiment of the collection of human ratings in the ANSS-70 dataset that described in section 5.2.3. Each of the 60 participants was supplied with an envelope having 70 cards with three sheets which included: an instruction sheet to collect human judgments, a sheet to record the similarity judgments and a sheet for personal information. The order of 70 cards was randomized before presentation to reduce the ordering effects.

5.3.3.3 Procedure

The same procedure was followed as in the ANSS-70 dataset to collect human ratings. The participants were asked to sort the cards into four groups in accordance with the similarity of meaning. The HSM group contained verb pairs between strongly related and identical in meaning. The High MSM group contained verb pairs which were very much alike in meaning, whilst the Low MSM group contained verb pairs which were vaguely similar in meaning and the LSM contained verb pairs unrelated in meaning. After sorting the cards, the participants were asked to check them carefully and then rank each verb pair using a point on a rating scale described by the semantic anchors which ran from 0.0 (unrelated in meaning) to 4.0 (identical in meaning). The instruction sheet also included some notes which enabled participants assigning an accurate degree of similarity by means of use of the first decimal place and to avoid using values lower than 0.0 or greater than 4.0 to rate the verb pairs.

5.3.3.4 Experimental Results and Discussion

The human similarity ratings collected in this experiment were calculated as the mean of the judgements provided by the 60 Arabic native speakers for each pair of verbs. Table 5.23 represents the results of this experiment which contains the set of 70 Arabic verb pairs with human ratings of similarity. The second and last pairs of columns represent the set of Arabic verb pairs in English and Arabic. The third column contains the mean of the similarity ratings collected from 60 Arabic native speakers whilst the fourth column represents the Standard Deviation (SD) of each verb pair.

The dataset is well balanced, if one considers that ~ 1/3 of the verb pairs are high, ~ 1/3 low and ~ 1/3 across the broad, difficult medium similarity band from 1.0 - 3.0. Figure 5.13 shows the distribution of the similarity ratings in the full AVSS-70 datasets.

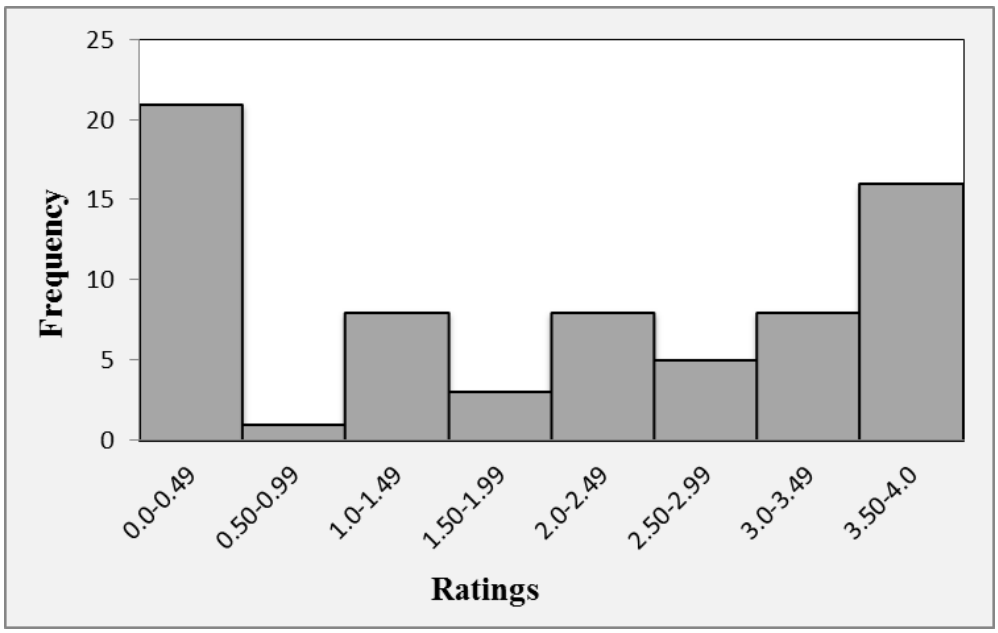


Figure 5.13 Distribution of the similarity ratings in AVSS-70 dataset.

Table 5.23 Arabic Verb Benchmark Dataset

	Verb Pairs		Human Ratings	SD	أزواج الافعـال	
1	Be superior	Depart	0.0	0	غـادر	تفـوق
2	Become	Be far	0.0	0.03	ابتعـد	اصبح
3	Be forced	Enrich	0.0	0	اغنـى	اضطـر
4	Waste	Explain	0.02	0.09	اوضح	هــدر
5	Continue	Buy	0.03	0.18	اشترى	استمر
6	Get	Seep	0.03	0.18	تسرب	نــال
7	Try	Be far	0.03	0.19	ابتعد	حـاول
8	Become	Contact	0.05	0.22	اتصل	اصبح
9	Enrich	Meet	0.05	0.22	الـتقى	اثرى
10	Include	Run across	0.06	0.23	قـابل	تضمن
11	Overcrowd	Wish	0.06	0.23	رغـب	اكتـظ
12	Rise	Understand	0.06	0.23	تفـهم	ارتفـع
13	Leak	Be obligatory	0.07	0.23	وجـب	نضـح
14	Find	Permit	0.08	0.36	اجــاز	لقــي
15	Require	Rise	0.09	0.34	ارتفـع	تطلـب
16	Ponder	Load	0.11	0.35	عبـأ	تأمـل
17	Appear	Grant	0.12	0.31	منـح	بـدا
18	Fill	Declare	0.14	0.50	صــرح	مــلأ
19	To be	Come	0.16	0.44	جــاء	صــار
20	Go	Believe	0.19	0.56	اعتقـد	ذهـب
21	Be capable	Comprise	0.21	0.60	شمـل	تمكـن
22	Happen	Find	0.79	0.72	وجـد	حدث
23	Find	Take	1.03	0.93	اخـذ	وجـد
24	Include	Exist	1.05	0.97	وُجِد	تضمن
25	Crowd	Restrict	1.06	0.97	حصـر	ازدحـم
26	Continue	Work	1.07	0.77	اشتغل	استمر
27	Give	Buy	1.11	0.89	اشتــرى	اعطــى
28	Enrich	Be strong	1.17	0.90	قــوي	اغنــى
29	Rise	Richen	1.20	1.01	اغنــى	ارتفـع
30	Consider	Ponder	1.33	1.05	تأمـل	اعتبـر
31	Appear	Find	1.60	1.12	وجـد	ظهـر
32	Contact	Run across	1.71	0.87	قابـل	اتصل
33	Restrict	Appoint	1.90	1.33	عيـن	حصــر
34	Obtain	Realize	2.00	1.17	ادرك	حصــل
35	Be capable	Be superior	2.11	1.11	تفـوق	تمكـن
36	Fill	Abound	2.12	1.06	كثـر	مــلأ
37	Try	Want	2.22	1	اراد	حــاول
38	Leak	Waste	2.25	1.09	هـدر	تسرب
39	Arrive	Catch up with	2.28	1.12	ادرك	وصــل
40	Hope	Request	2.36	1	طلـب	تمنــى
41	Appoint	Specify	2.46	1.06	ثبـت	عيـن
42	Be forced	Be obligatory	2.52	1.28	وجـب	اضطـر
43	Declare	Explain	2.72	1.03	اوضـح	صــرح
44	Depart	Be far	2.75	1	ابتـعد	غــادر

45	Approve	Understand	2.98	0.91	تفهم	وافق
46	Be obligatory	Require	2.98	0.99	تطلب	وجب
47	Contact	Meet	3.00	0.77	التقى	اتصل
48	Leak	Seep	3.06	1.17	تسرب	نضح
49	Believe	Consider	3.07	0.88	اعتقد	اعتبر
50	Increase	Rise	3.11	0.85	ارتفع	ازداد
51	Happen	Take place	3.22	0.83	جرى	حدث
52	Arrive	Come	3.41	0.87	جاء	وصل
53	Try	Endeavour	3.42	0.74	سعى	حاول
54	Appear	Appear	3.44	0.75	ظهر	بدا
55	Include	Comprise	3.50	0.83	شمل	تضمن
56	To be	Become	3.51	0.83	اصبح	صار
57	Enrich	Richen	3.53	1.01	اغنى	اثرى
58	Find	Find	3.55	0.81	وجد	لقي
59	Appoint	Employ	3.63	0.83	وظف	عين
60	Go	Depart	3.66	0.59	غادر	ذهب
61	Be capable	Be able	3.68	0.72	استمكن	تمكن
62	Hope	Wish	3.69	0.55	رغب	تمنى
63	Allow	Permit	3.75	0.51	اجاز	سمح
64	Fill	Load	3.78	0.48	عبأ	ملأ
65	Announce	Declare	3.79	0.57	صرح	اعلن
66	Continue	Go on	3.85	0.39	واصل	استمر
67	Accept	Approve	3.86	0.39	وافق	قبل
68	Give	Grant	3.87	0.37	منح	اعطى
69	Get	Obtain	3.87	0.39	حصل	نال
70	Overcrowd	Crowed	3.88	0.32	ازدحم	اكتظ

The ratio scale had been identified in the noun dataset creation procedure as a suitable measurement scale used for both word semantic similarity measures and datasets. In addition, the correlation coefficient has been considered as a suitable statistic that can be applied for measures made on a ratio scale. Consequently, the Pearson product moment correlation coefficient was used in this dataset to identify the consistency of similarity judgements for each participant with the rest of group using the leave-one-out resampling technique as described in section 5.2.4. For each of 60 participants, the correlation coefficient was calculated between the participant's ratings and the average ratings of the rest of group. Figure 5.14 shows the consistency of the similarity judgements of the 60 participants.

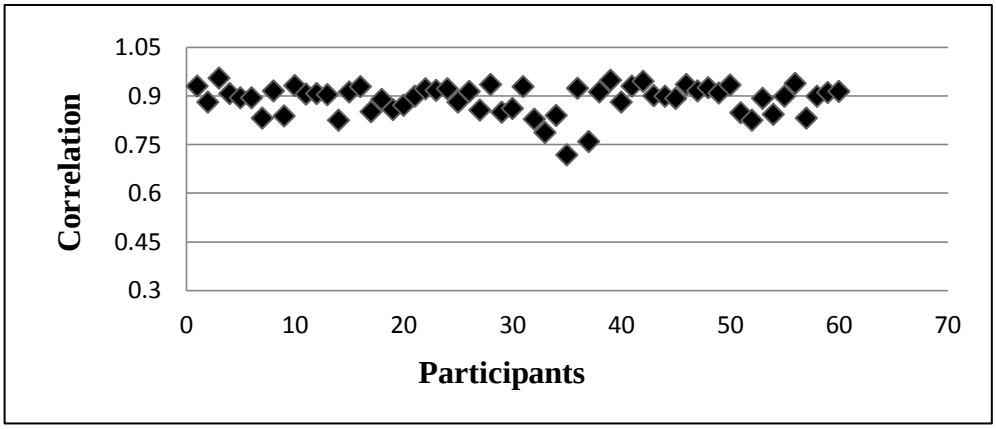


Figure 5.14 The Correlation Coefficients of 60 Arabic Participants

The possible indicative value and bounds of performance expected from a computational Arabic verb similarity algorithm attempt to perform the same task have been calculated as the average, worst and best performances of human participants on the AVSS-70 dataset as shown in table 5.24.

Table 5.24 Correlation Coefficient with Mean Human Judgements

	Correlation r
Average of the correlation of all participants	0.888
Best participants	0.954
Worst participants	0.718

Any Arabic verb similarity measure that equals or exceeds the average of the correlations of all participants (0.888) can be considered performing well. The worst performing participant of 0.718 is considered as the lower bound for the expected performance whilst any verb similarity measure coming close to the best performing participant at 0.954 would be considered as performing very well.

5.3.4 The Evaluation Procedure

The development process of the Arabic verb similarity (KalTa-F) measure (described in chapter 4) consisted of two steps. The first step involved creating a measure that calculated the verb similarity using information sources extracted from the verb

hierarchy in AWN. A hybrid approach was presented in the second step to identify the similarity rating between two Arabic verbs based on their roots and using information sources extracted from the verb hierarchy and noun hierarchy in AWN. For the purpose of comparison, the verb measure created in the first step was called as KalTa-F without Root whilst the second step's measure was called as KalTa-F. These measures were evaluated in accordance with the same procedure used to evaluate the KalTa-A measure which required:

1. Partitioning the AVSS-70 dataset into two sub-datasets.
2. Identifying the optimal parameters values (α and β) for each verb measure.
3. Validation of the KalTa-F without Root and KalTa-F measures using the optimal parameter values.

The role of α and β parameters was explored by partitioning the AVSS-70 dataset into two sets known as training and evaluation datasets. The training dataset was used to search the suitable parameters α and β within the interval [0, 1] whilst the evaluation dataset was employed to identify the KalTa-F without Root and KalTa-F algorithms validated. Each dataset consisted of 35 verb pairs spanning the similarity of meaning range from maximum to minimum, which were selected as follows.

1. The original AVSS-70 dataset consisted of 22 low similarity, 24 medium similarity and 24 high similarity verb pairs. Therefore, each sub-dataset contained 11 low similarity, 12 medium similarity and 12 high similarity verb pairs.

2. For each similarity class within the same sub-dataset, the verb pairs were selected with similarity of meaning ranging from low to high. Figures 5.15 and 5.16 present the verb pairs in the evaluation dataset and training dataset respectively.

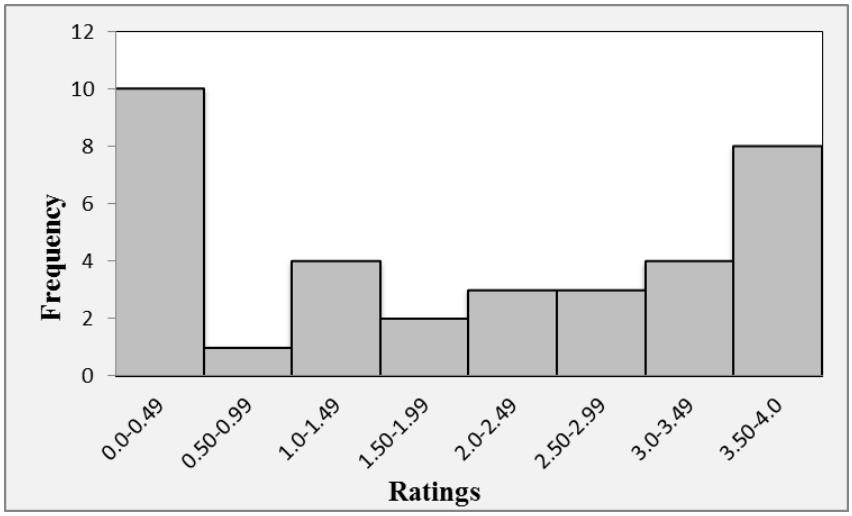


Figure 5.15 Distribution of similarity ratings in the evaluation dataset

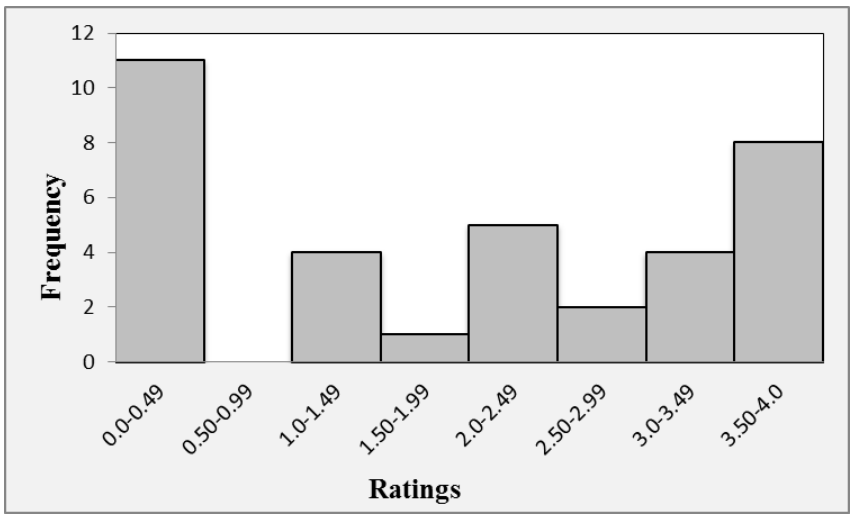


Figure 5.16 Distribution of similarity ratings in the training dataset

Some Arabic words have been not added to the current version of AWN consequently only 30 verb pairs of each sub-datasets have been used in this study's experiments. The verb pairs in the training and the evaluation datasets are listed with human ratings in table 5.25 and table 5.26, respectively. The faint grey verb pairs have not been used in the training and evaluation experiments.

Table 5.25 Training Dataset Verb Pairs with Human Ratings

No.	Verb Pairs		Human Ratings	أزواج الافعـال	
1	Be forced	Enrich	0.0	اغنـى	اضطـر
2	Try	Be far	0.01	ابتعد	حـاول
3	Get	Seep	0.01	تسـرب	نـال
4	Become	Contact	0.01	اتصل	اصبح
5	Rise	Understand	0.01	تفهـم	ارتفـع
6	Include	Run across	0.01	قـابل	تضمن
7	Leak	Be obligatory	0.02	وجـب	نضـح
8	Find	Permit	0.02	اجـاز	لقـي
9	Ponder	Load	0.03	عبـأ	تأمـل
10	Fill	Declare	0.04	صـرح	مـلأ
11	Be capable	Comprise	0.05	شمـل	تمكـن
12	Include	Exist	0.26	وجد	تضمن
13	Continue	Work	0.27	اشتغل	استمر
14	Rise	Richen	0.30	اغنـى	ارتفـع
15	Consider	Ponder	0.33	تأمـل	اعتبـر
16	Appear	Find	0.40	وجـد	ظهـر
17	Obtain	Realize	0.50	ادرك	حصـل
18	Fill	Abound	0.53	كثـر	مـلأ
19	Leak	Waste	0.56	هـدر	تسـرب
20	Hope	Request	0.59	طلـب	تمنـى
21	Appoint	Specify	0.61	ثبـت	عيـن
22	Depart	Be far	0.69	ابتـعد	غـادر
23	Be obligatory	Require	0.75	تطلـب	وجـب
24	Contact	Meet	0.75	التقـى	اتصـل
25	Believe	Consider	0.77	اعتقـد	اعتبـر
26	Happen	Take place	0.80	جـرى	حـدث
27	Appear	Appear	0.86	ظهـر	بـدا
28	To be	Become	0.88	اصـبح	صـار
29	Enrich	Richen	0.88	اغنـى	اثرى
30	Go	Depart	0.91	غـادر	ذهـب
31	Be capable	Be able	0.92	استمكـن	تمكـن
32	Allow	Permit	0.94	اجـاز	سمـح
33	Announce	Declare	0.95	صـرح	اعلـن
34	Accept	Approve	0.96	وافـق	قبـل
35	Give	Grant	0.97	منـح	اعطـى

5.3.4.1 Tuning Parameters

Given the initial value of each parameter (α and β), the verb pairs on the training dataset were run using the KalTa-F (with Root) measure to produce the machine similarity ratings in a range from 0 to 1. The correlation coefficient between the human ratings and those obtained from the KalTa-F measure was calculated. The values of the Arabic measure parameters were changed to obtain a set of correlation coefficients. The increasing step of α and β was 0.05. The parameters with the strongest correlation coefficient were considered as the optimal parameters. For the KalTa-F measure, the strongest correlation coefficient was obtained at $\alpha = 0.2$ and $\beta = 0.459$.

The optimal parameters values of the KalTa-F without Root measure were identified following the same procedure used to obtain the optimal values of the KalTa-F algorithm parameters. The strongest correlation coefficient was obtained at $\alpha = 0.35$ and $\beta = 0.96$.

Using the identified optimal parameters of each measure, the verb pairs on the evaluation dataset were run to generate the machine similarity ratings to assess the accuracy of KalTa-F without Root and KalTa-F measures. Table 5.26 shows the human similarity ratings with the corresponding machine similarity ratings on the evaluation dataset. The first and last columns represent the verb pairs on the evaluation dataset whilst the second column represents the human similarity ratings which were rescaled from 0 - 4 to 0 – 1 for the purpose of comparison. The third and fourth columns represent the corresponding machine similarity ratings produced by the KalTa-F without Root and KalTa-F measures respectively. The validation of each algorithm was identified by calculating the correlation coefficient between the human ratings and the ratings obtained from each measure on the evaluation dataset.

Table 5.26 The Evaluation Dataset Verb Pairs with Human and Machine Ratings

No.	Verb Pairs		Human Ratings	KalTa-F without Root	KalTa-F Ratings	أزواج الافعـال	
1	Be superior	Depart	0.0	0	0.13	غـادر	تفـوق
2	Become	Be far	0.0	0	0	ابتعـد	اصبح
3	Waste	Explain	0.01	0	0	اوضح	هــدر
4	Continue	Buy	0.01	0.26	0.22	اشترى	استمر
5	Enrich	Meet	0.01	-	-	الـتـقى	اثرى
6	Overcrowd	Wish	0.01	0	0	رغـب	اكتـظ
7	Require	Rise	0.02	0	0.19	ارتفـع	تطلـب
8	Appear	Grant	0.03	0	0.16	منـح	بـدا
9	To be	Come	0.04	0	0	جـاء	صــار
10	Go	Believe	0.05	0	0.14	اعتقـد	ذهـب
11	Happen	Find	0.20	0	0.49	وجـد	حدث
12	Find	Take	0.26	0.37	0.72	اخـذ	وجـد
13	Crowd	Restrict	0.26	-	-	حصـر	ازدحـم
14	Give	Buy	0.28	0	0.52	اشترى	اعطـى
15	Enrich	Be strong	0.29	0.35	0.23	قــوي	اغنــى
16	Contact	Run across	0.43	0	0.59	قابـل	اتصـل
17	Restrict	Appoint	0.48	0	0.78	عيـن	حصــر
18	Be capable	Be superior	0.53	0	0.19	تفـوق	تمكـن
19	Try	Want	0.56	0	0.49	اراد	حــاول
20	Arrive	Catch up with	0.57	0	0.59	ادرك	وصـل
21	Be forced	Be obligatory	0.63	-	-	وجـب	اضظـر
22	Declare	Explain	0.68	0.35	0.52	اوضـح	صــرح
23	Approve	Understand	0.75	0	0.87	تفهـم	وافـق
24	Leak	Seep	0.76	-	-	تسرب	نضـح
25	Increase	Rise	0.78	0.96	0.99	ارتفـع	ازداد
26	Arrive	Come	0.85	0	0.63	جـاء	وصـل
27	Try	Endeavour	0.86	0.96	0.88	سعـى	حــاول
28	Include	Comprise	0.88	0.96	0.81	شمـل	تضمـن
29	Find	Find	0.89	0.96	0.98	وجـد	لقــي
30	Appoint	Employ	0.91	0.74	0.99	وظـف	عيـن
31	Hope	Wish	0.92	0.96	0.95	رغـب	تمنــى
32	Fill	Load	0.95	0.99	0.99	عبــأ	مــلأ
33	Continue	Go on	0.96	1	0.98	واصـل	استمـر
34	Get	Obtain	0.97	0.37	0.99	حصــل	نــال
35	Overcrowd	Crowed	0.97	-	-	ازدحــم	اكتــظ

5.3.5 Findings and Discussion

The possible indicative value and bounds of performance expected from a computation Arabic verb similarity algorithm attempting to perform the same task have been calculated as the average, worst and best performances of human

participants on the evaluation dataset as shown in table 5.27. This was carried out by means of the leave-one-out resampling technique to calculate the correlation coefficient of each of the 60 participants with the rest of the group.

The consistency of each verb measure (KalTa-F without Root and KalTa-F) with human perception was identified by computing the correlation coefficient between the average ratings of human participants on the evaluation dataset and the machine ratings obtained from each verb measure as shown in Table 5.27.

The KalTa-F (with Root) measure achieved a good value of the Pearson correlation coefficient (r = 0.906) with the human judgments as shown in figure 5.17. The KalTa-F measure performed very well at (r= 0.906) with the average value of the correlations of human participants (r = 0.887). Furthermore, the performance of the KlaTa-F measure was substantially better than the worst human (lower bound) performance at (r=0.745).

Table 5.27 Performance of KalTa-F without Root and KalTa-F Measures on the Evaluation Dataset

On the Evaluation Dataset	Correlation r
KalTa-F without Root measure	0.715
KalTa-F measure	0.906
Average of the correlation of all participants	0.887
Best participants	0.961
Worst participants	0.745

Table 5.27 shows that the KalTa-F without Root measure achieved a correlation significantly below the average of the correlation of human performance. The result from a one sample t-test which was used to compare a single correlation (KalTa-F without Root) with the average of the correlation coefficients on the evaluation dataset.

Null hypothesis (H_0) is a test of $\mu = 0.715$ vs. $\mu \neq 0.715$. The result of the one sample t-test with confidence interval plot is summarized in appendix 4. The true mean could lie anywhere in the interval (0.875, 0.899), the sample mean (n=60) is

0.887 and t-test statistic is 28.75 with P-value < 0.0001. Since the P-value is less than the significance level at ($\alpha = 0.05$ and $\alpha = 0.01$), the null hypothesis can reject.

The KalTa-F without Root measure also achieved a correlation coefficient (r = 0.715) lower than the worst participant correlation (r=0.745) due to the limitations of the verb hierarchy as described in chapter 4. Figure 5.18 illustrates the correlation coefficient between the ratings obtained from the KalTa-F without Root measure and the ratings provided by humans. As shown in figure 5.18 and table 5.26, the majority of verb pairs rated medium by participants achieved very low machine similarity values which were equal to 0. Also some verb pairs rated high by participants attained very low or medium similarity values using KalTa-F without Root measure, for instance verb pair numbers 23, 26, 30 and 34.

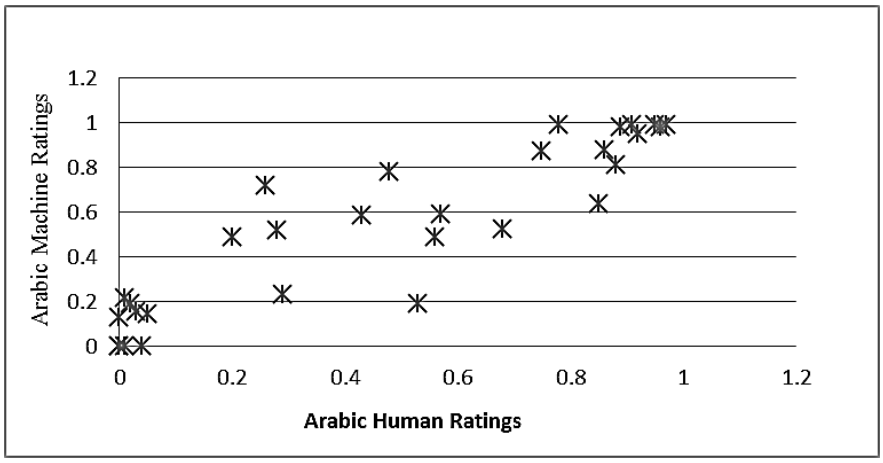


Figure 5.17 The Correlation Coefficient between Human Ratings and KalTa-F Measure Ratings on the Evaluation Dataset

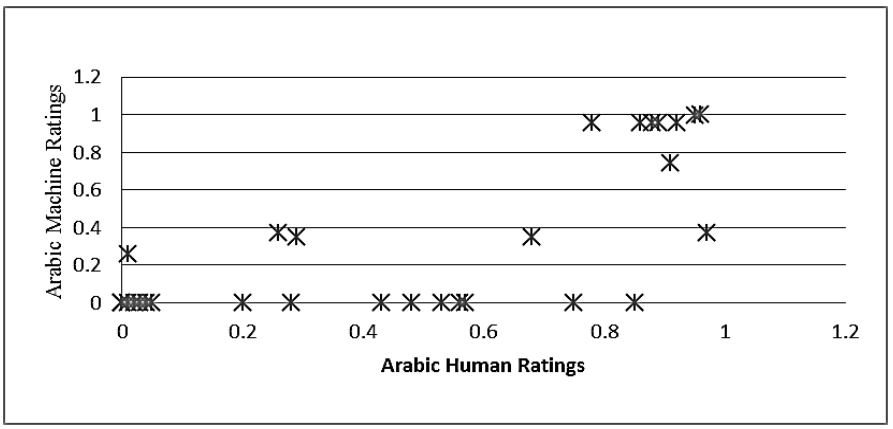


Figure 5.18 The Correlation Coefficient between Human Ratings and KalTa-F without Root Measure Ratings on the Evaluation Dataset

Steiger's z-test was used to compare the difference between KalTa-F measure and KalTa-F without Root measure. Using Steiger's z-test requires the construction of a correlation triangle (3 correlations) between:

KalTa-F without Root ratings vs. Human ratings = 0.715
KalTa-F ratings vs. Human ratings = 0.906
KalTa-F without Root vs. KalTa-F = 0.725

n = 30 (the number of verb pairs in the evaluation dataset)

Applying the Steiger's z-test (using the online calculator which was available at (Grabin, 2013)) indicates that the difference between KalTa-F and KalTa-F without Root measures is statistically significant (Z= -2.84, p = 0.004).

5.4 Conclusions

This chapter has described the production of the first two Arabic word benchmark datasets, the Arabic noun similarity dataset (ANSS-70) and the Arabic verb similarity dataset (AVSS-70). These datasets will make a substantial contribution to future work in the field of Arabic word semantic similarity and should be considered as a reference basis from which to evaluate and compare developing methodologies from researchers in the field.

Though it is not possible to cover the language comprehensively with a delimited number of word pairs (70 pairs) in each dataset, this research used a systematic process to select the set of Arabic stimulus nouns and the set of Arabic stimulus verbs. In the noun (ANSS-70) dataset, a new method was used to select the stimulus nouns by means of the creation of 27 Arabic categories with 27 different themes to promote the best possible semantic representation. As regards the Arabic verb dataset, the sampling frame technique was used to choose the stimulus verbs by decomposing the Arabic verbs into a hierarchy of classes based on established grammatical techniques developed for Arabic NLP.

Unlike the prior work, participants were chosen to produce a set of 70 noun pairs for the ANSS-70 dataset and a set of 70 verb pairs for the AVSS-70 dataset which covered a range of semantic similarity values from maximum to minimum. Human ratings were collected for each dataset using the best possible available techniques.

The samples of participants used in the ANSS-70 and AVSS-70 datasets experiments were selected to achieve a balance and also representation of the human population well beyond that of prior work. Furthermore, the procedures used for production of these datasets can be used by other Arabic researchers to extend the ANSS-70 and AVSS-70 datasets.

The motivation for the creation of these datasets was to identify the validation of the KalTa-A and KalTa-F measures before integrating them into the ASTSS measure. This chapter described the evaluation procedure of each measure which involved the optimization of parameters in the algorithm by means of partitioning the dataset into training and evaluation sets. Experimental evaluation of the KalTa-A measure indicated that the use of SUMO improved the performance of KalTa-A measure which achieved a good correlation compared with the average value of human participants. This measure with its optimal parameter values ($\alpha = 0.12$ and $\beta = 0.21$) will be used with the ASTSS measure.

Furthermore, experimental evaluation of the Arabic verb measure showed that the performance of KalTa-F measure is significantly better than the KalTa-F without Root measure performance. Consequently, KalTa-F (with Root) with its optimal parameters values ($\alpha = 0.2$ and $\beta = 0.459$) will be used with the ASTSS measure.

Chapter 6

Arabic Short Text Semantic Similarity Measure Evaluation

6.1 Introduction

The evaluation of the new Arabic short text semantic similarity framework, namely that of NasTa, presented in chapter 4 will be described in this chapter which comprises the following steps:

1. The production of an Arabic Short Text benchmark (ASTSS-68) dataset.
2. The creation of an optimization short text pairs set (ASTSS-21).
3. The procedure used in the evaluation of the NasTa-A algorithm created in the first phase of the NasTa framework development process which is based on the noun semantic similarity and word order similarity.
4. The evaluation procedure for the NasTa-F algorithm created in the second phase of the NasTa development process which is based on the part of speech and word sense disambiguation.

The ASTSS-68 dataset is designed to meet the three issues of the dataset design process highlighted in chapter 3. Firstly, selection of a sample of the short text pairs that represents the properties of the Arabic language. The produced dataset consists of 68 Arabic short text pairs which are generated using a range of resources from traditional Arabic grammar to grammatical techniques developed for Arabic NLP. Secondly, collection of similarity ratings that precisely represent the human perception of similarity using a representative sample of participants. Human ratings are collected using the best possible available techniques as identified in chapter 5. Thirdly, determination of the appropriate statistical measures that can be applied to make judgements about the short text similarity measures. The correlation coefficient (considered in the noun and verb datasets creation procedures as a most suitable) is used for reporting the ASTSS-68 dataset experimental results.

The optimization set (ASTSS-21) is used to determine the optimal parameter values of the NasTa which is the most important step in the evaluation process of the NasTa-A and the NasTa-F algorithms. The process of the optimization of parameters

in the algorithms and the procedures used to assess the accuracy of the NasTas-A and NasTa-F will be described in this chapter.

6.2 The Arabic Short Text Benchmark Dataset (ASTSS-68)

The methodology used to create the first short text benchmark dataset for MSA, namely that of ASTSS-68 is presented in this section. The ASTSS-68 dataset design process consists of four stages which include:

1. Selection of the stimulus words.
2. Production of a database of Arabic short texts.
3. Selection of 68 pairs of Arabic short texts from the database.
4. Collection of the human similarity ratings for 68 short text pairs.

The ASTSS-68 dataset adapted elements from the work of the English text semantic similarity (O'Shea et al., 2013) to select Arabic stimulus words and create the short text database taking into account the Arabic language features described in chapter 2. The procedure of collection of the human similarity ratings is adopted from the work of Arabic Noun (ANSS-70) dataset (chapter 5, section 5.2.3).

6.2.1 Selection of the Stimulus Words

Representation of the Arabic language was achieved by carefully selecting a set of stimulus words by means of adoption of a sampling frame technique used by (O'Shea et al., 2013). This technique was used in the Arabic Verb (AVSS-70) benchmark dataset (chapter 5) to select the set of stimulus verbs and is expanded in this section to select the set of the stimulus words which is comprised of nouns, verbs, adjectives and adverbs. A sampling frame is a method of representing a large population with a small carefully-chosen sample randomly selected with constraints. The size of the Arabic stimulus words set chosen to create ASTSS-68 was 64 which was selected based on the principals of sampling frame (O'Shea et al., 2013) plus 4 words to represent specific Arabic features described later in section (6.2.1.1). The selection process consisted of two steps:

1. Decomposing the Arabic words into a hierarchy of classes.
2. Population of the slots in the frame with Arabic words.

6.2.1.1 Decomposing the Arabic Words into a Hierarchy of Classes

In this study, the Arabic words were decomposed into a tree structure based on special syntactic and semantic features. Each of the tree levels is described in this section.

Traditional and modern Arabic linguistics classified Arabic words into 2 classes useful in a top-level decomposition which are content words and function words. However, they differed in the classification of content words and function words as described in chapter 2. Traditional linguistics and current Arabic grammar books classified the content words into nouns and verbs only. Whereas, modern linguistics considered this classification insufficient for a highly inflectional language such as Arabic and they classified the content words into nouns, verbs, adjectives and adverbs. This research followed the modern classification of Arabic words. The content words were decomposed into nouns, verbs, adjectives and adverbs which were useful in second-level decomposition as shown in figure 6.1. The function words based on the modern classification method included (preposition, pronouns, articles, etc.) which naturally appear in the short text. Consequently, only the content words were included in the sampling frame.

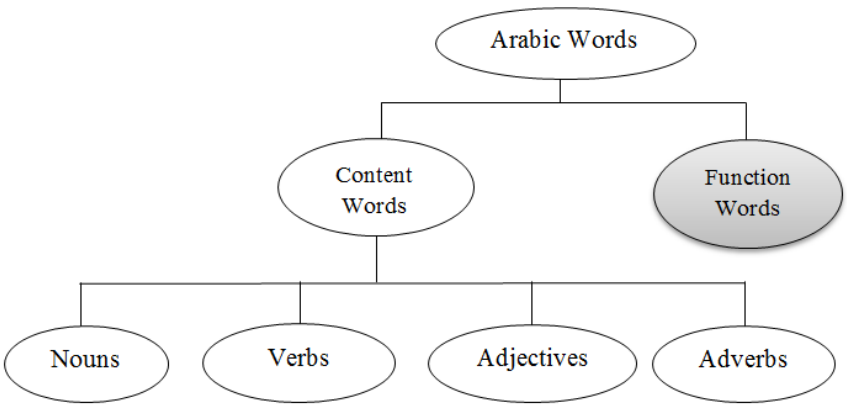


Figure 6.1 The top and second levels Arabic word decomposition

Each of the content word classes was further decomposed in the next stage of the Arabic word decomposition process. However, performing this required first the determination of the proportion of occurrence of each class (nouns, verbs, adjectives and adverbs) in the final set of Arabic stimulus words. This was achieved by using the Arabic Word Count (AWC) corpus (Attia et al., 2011) based on the assumption that the high frequency words should have a higher probability of appearing in the sample frame. The total number of occurrences of all of words in each content class in the AWC frequency list was calculated. This is then used to determine the proportions of occurrence of each content class in the final set of Arabic stimulus words. The size of the Arabic stimulus words set is 64 which was established as a balance of size and effort and selected based on the principals of sampling frame (O'Shea et al., 2013). The distribution of 64 Arabic stimulus words between the content words classes using the AWC list is presented in table 6.1.a. The results in table 6.1.a indicate that more than half of the stimulus words were allocated to Arabic nouns whilst a very limited number were allocated to adverbs. This follows the traditional classification of Arabic words. However, Rydin (2005) reported that "In Arabic, few words are adverbs in and of themselves; but there are some (such as ***faqaT*** فقط 'only') and most words that function as Arabic adverbs are nouns in the accusative case". Consequently, it was decided to take 8 slots from the nouns and allocate them to adverbs. Also, some of the Arabic adjectives are nouns in the accusative case therefore the number of the stimulus words was increased to 68 and the additional four words were allocated to adjectives. The final distribution of 68 stimulus words between the content words classes is presented in table 6.1.b.

Table 6.1 Distribution of the Arabic stimulus words between the content words classes.

(a)

64 Stimulus Words	Content Words Classes
38	Nouns
15	Verbs
10	Adjectives
1	Adverbs

(b)

68 Stimulus Words	Content Words Classes
30	Nouns
15	Verbs
14	Adjectives
9	Adverbs

Each of the content words classes is decomposed further in the following sections.

Decomposition of the Arabic Nouns

Abdul-Raof (2000) decomposed Arabic nouns into concrete nouns such as ***mother*** and abstract nouns such as ***government***. This offered a top level of noun decomposition as shown in figure 6.2 which shows the Arabic noun sub-tree structure. The abstract nouns were decomposed further into two classes by (Abdul-Raof, 2000) which included human and inanimate. Examples of human class include انسانية "humanitarian", ديانة "religion", عادات "habits", زواج "marriage", etc. An inanimate class was decomposed into 6 sub-classes at the low level which are: fact, place, action, time, mental and emotion as shown in figure 6.2. Consequently, the final decomposition of the abstract nouns consisted of 7 classes including the human and 6 inanimate sub-classes.

As shown in table 6.1.b, 30 slots were reserved for Arabic nouns. Based on observation of examples listed in (Abdul-Raof, 2000), it was decided to allocate 7 slots for abstract nouns. 21 slots of the remainder were allocated for concrete nouns and 2 allocated for special language properties which are discussed later.

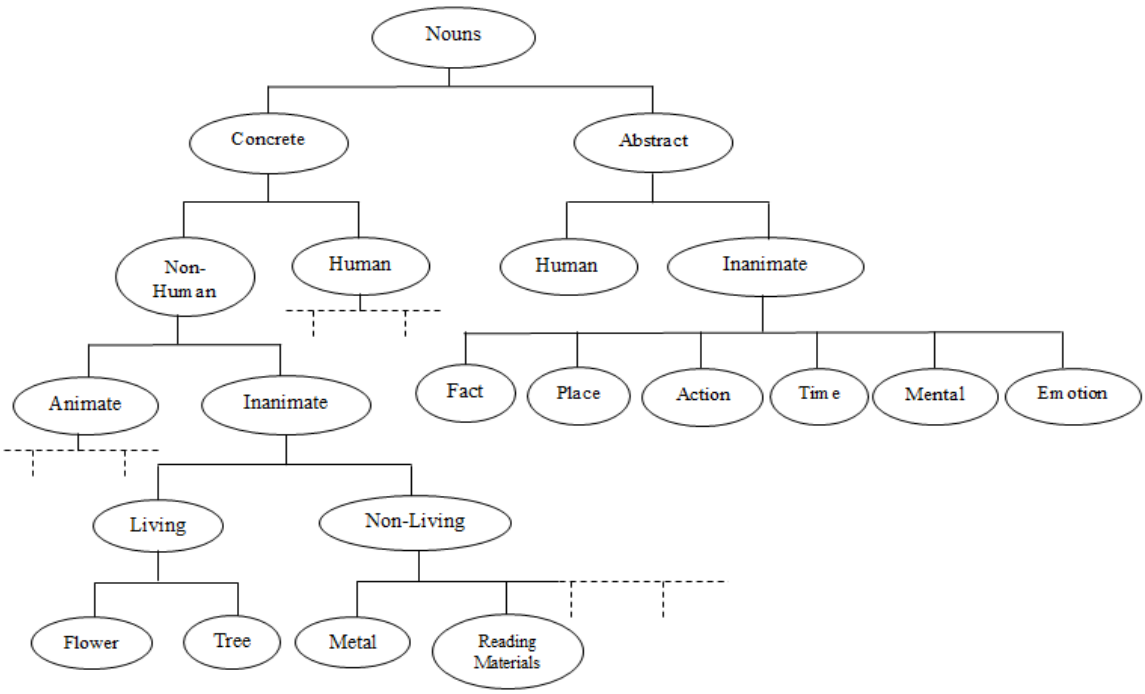


Figure 6.2 Arabic nouns sub-tree structure

The final allocation of 7 abstract nouns slots involved allocation of one slot to each of 7 abstract noun classes which include human, inanimate fact, inanimate place, inanimate action, inanimate time, inanimate mental and emotion.

Abdul-Raof (2000) also decomposed the Arabic concrete nouns into human and non-human. The non-human concrete nouns were decomposed further into animate such as ***animals*** and inanimate such as ***book*** or ***tree***. Some studies considered for example the fruit, vegetables, trees and flowers as living (O'Shea, 2010) because they are alive but they do not move. The inanimate class was used in some English work with both living and non-living categories as reported by (Caramazza and Shelton, 1998). Therefore, the inanimate class was decomposed further into living such as ***tree*** and non-living such as ***book*** as shown in figure 6.2.

The Arabic category norms created in the ANSS-70 dataset such as (family members, birds) were used in the fine-grained decomposition of concrete nouns, which were held to be a good source of semantic categories. Chapter 2 highlighted that the content of the category norms differ from one language to another on the basis of the culture (Yoon et al., 2004). Consequently, The 27 Arabic categories (described in chapter 5 section 5.2.1) were extended to include 20 additional categories created form 20 themes taken from English category norms. These were needed to support the decomposition of the animate and inanimate/living categories and also to promote the semantic representation for other concrete nouns' classes. Appendix 5 contains a list of new categories created in this dataset. The distribution of 47 Arabic categories between the concrete nouns classes is presented in table 6.2. Consequently, the final stage of the concrete nouns decomposition used categories from the 47 Arabic category norms created in this research as shown in figure 6.2.

Table 6.2 the distribution of Arabic categories between the concrete nouns classes

Concrete nouns classes	**47 Arabic categories**
Human	10
Non-Human/animate	6
Non-Human/Inanimate/Living	4
Non-Human/Inanimate/Non-Living	27

As mentioned earlier, 21slots were allocated for concrete nouns. 6 slots would be allocated for concrete human, 6 slots allocated for non-human/animate, 7 slots

allocated for non-human/inanimate/non-living and 2 allocated for non-human /inanimate/living (has relatively minor role). Table 6.3 shows the Arabic categories selected for each concrete noun class. For each concrete noun class, the categories were selected randomly from the categories allocated to this class as in table 6.2.

Table 6.3 the final allocation of concrete noun slots

No	Concrete Nouns classes	Arabic Category Selected
1	Human	Family members
2	Human	Military title
3	Human	An occupation
4	Human	wise person
5	Human	part of human body
6	Human	type of male's life stages
7	Non-Human /Animate	Four footed animals
8	Non-Human /Animate	Birds
9	Non-Human /Animate	Insect
10	Non-Human /Animate	Fish
11	Non-Human /Animate	Snake
12	Non-Human /Animate	Diseases
13	Non-Human/Inanimate /Living	Tree
14	Non-Human/Inanimate /Living	Flower
15	Non-Human/Inanimate /Non-Living	Type of reading material
16	Non-Human/Inanimate /Non-Living	Building for religious services
17	Non-Human/Inanimate /Non-Living	Weapon
18	Non-Human/Inanimate /Non-Living	Weather phenomenon
19	Non-Human/Inanimate /Non-Living	Transportation vehicle
20	Non-Human/Inanimate /Non-Living	Non-alcoholic beverage
21	Non-Human/Inanimate /Non-Living	Part of day

Additional Features

Certain features (linguistic features) of Arabic words such as polysemy and homophony may affect perceived similarity. The words are polysemous which means they have one spelling and pronunciation with multiple meanings. For example, the Arabic word جبن which mean cheese or cowardice. Like other natural languages, most Arabic words are polysemous, therefore some of them will be included

automatically in the sampling frame. Consequently, it was decided to eliminate this feature (polysemy) from needing representation.

Words are homographs which share the same spelling but different pronunciations, usually with different meaning. For example, the Arabic word بر could mean three different nouns, بَرّ *barr* "land" or بُرّ *burr* "wheat" and بِرّ *birr* "reverence". The homograph in the Arabic language results from missing diacritics in the contemporary Arabic writing system. It was decided to apply this feature to some content classes as a homograph pair. For example, the homograph noun-verb pair (e.g., ذهب as a noun "gold" or verb "go").

Moreover, it was decided to apply the oppositeness of meaning (antonymy feature) to some content word classes. Finally, the property of degree for the adjectives and adverbs is represented in the sampling frame. For example, the adjective واضح "clear" has the comparative اوضح "clearer".

Decomposition of the Arabic Verbs

The method of the creation of the Arabic verb (AVSS-70) dataset presented in chapter 5 decomposed the Arabic verbs into a tree structure using grammatical techniques developed for NLP which include Case Grammar (CG) (Al-Qahtani, 2005) and Arabic VerbNet (AVN) (Mousser, 2010). This method was used to decompose the Arabic verbs in the ASTSS-68 dataset.

Figure 6.3 shows a portion of the Arabic verb sub-tree structure. At the top level of the tree structure, the Arabic verbs were decomposed into 3 classes based on CG classification which are state, process and action. Each verb class was decomposed into basic, experiential, benefactive, and locative verbs at the intermediate level of Arabic verb hierarchy. These sub-classes were employed in the next stage decomposition of verbs.

Mousser (2010) presented a large coverage verb lexicon for the Arabic language which exploited Levin's verb-classes (Levin, 1993), as described in chapter5. This work offered good verb classes for Arabic which were used in the final stage decomposition in this study. Combining CG and AVN verb classes for

decomposition offered a good intermediate structure and fine-grained classes which were easy to understand and use. In the final level of Arabic verb decomposition, each CG verb class at the intermediate level was combined with a different class of the AVN verb classes from different level as shown in figure 6.3.

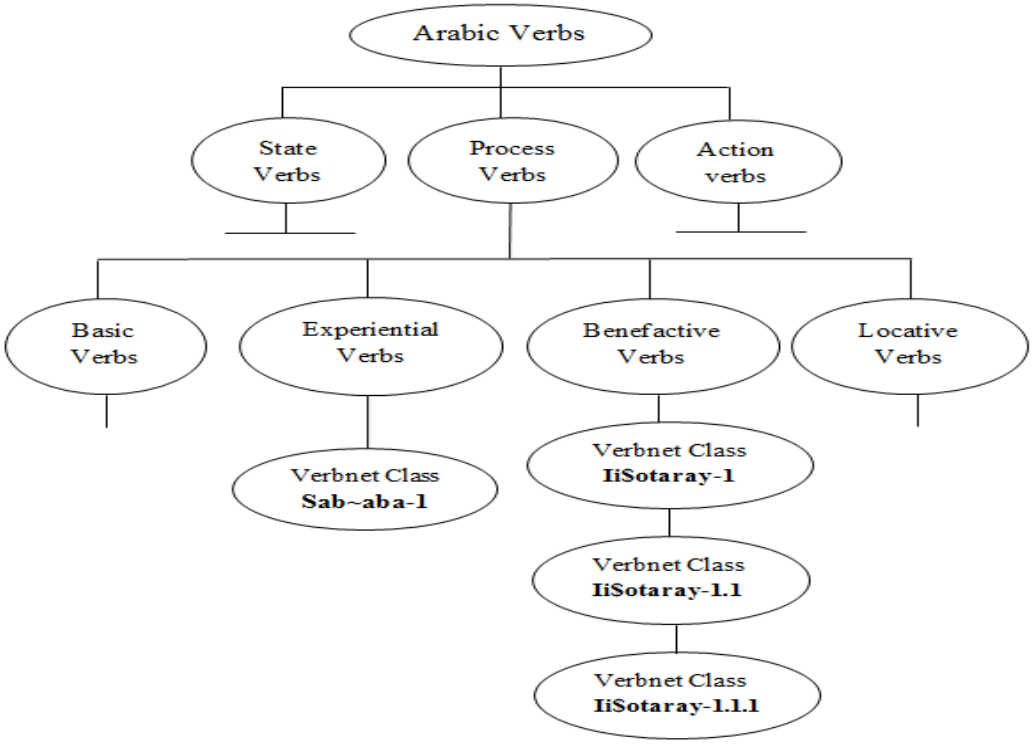


Figure 6.3 A portion of the Arabic verbs sub-tree

As presented in table 6.1.b, 15 slots were reserved for Arabic verbs and it was decided to allocate 12 slots to high-level verb classes whereby 4 slots were allocated to state verbs class, 4 slots were allocated to process verb class and 4 slots were allocated to action verb class. Within each of the high level verb classes, 1 slot was allocated to basic verb class, 1 slot to experiential verb class, 1 slot to benefactive verb class and 1 slot to locative verb class. Table 6.4 shows the full allocation.

Each of 12 verb classes was also allocated to a different AVN verb class, whereby 9 slots were allocated to the top level of the AVN verb classes, as shown in table 6.4. 3 slots (one from each high level CG classes including state, process and action) were allocated to the lower level of the AVN verb classes (second and third).

Table 6.4 Arabic verb sub-frame

Class	Verb Types	AVN Verb classes
1	State Basic	Class a
2	State Experiential	Class b
3	State Benefactive	Class c
4	State Locative	Class d
5	Process Basic	Class g
6	Process Experiential	Class h
7	Process Benefactive	Class i
8	Process Locative	Class p
9	Action Basic	Class q
10	Action Experiential	Class r
11	Action Benefactive	Class y
12	Action Locative	Class z
13	Any State verb	Paired with any AVN verb class
14	Any Process verb	Paired with any AVN verb class
15	Any Action verb	Paired with any AVN verb class

In order to promote the high semantic similarity, it was decided to allocate the remainder of 15 slots (3 slots) to verbs paired with one of the AVN classes already used, as shown in table 6.4.

To promote polysemy, the decision was made to ensure that at least one of the verbs chosen appeared in several AVN verb classes and also at least one verb appeared in only one AVN verb class.

As discussed under Arabic nouns decomposition, two constraints were appended that one of the verbs chosen must be part of a homograph noun-verb pair and also one verb must be part of a homograph verb-adjective pair.

Decomposition of the Arabic Adjectives

As mentioned previously, traditional linguistics and current Arabic grammar books classify content words into nouns and verbs only (Suleiman, 1990) where the nouns include adjectives and adverbs. Moreover, in the absence of current research on the resolution of categorical intersection between nouns and adjectives (Attia, 2008), it was decided to go back approximately eleven hundred years when the Arabic grammarian Ibn as-Sarraj (2009) in his book ***al-Usool fi an-Nahw*** distinguished five types of Arabic adjectives from the nouns. Consequently, the Arabic adjectives were decomposed into five classes as described by Ibn as-Siraaj. The five adjective classes

included a visible quality (***the Hilya***), an internal trait, an action, an adjective of relation (***Nasab***), and a descriptive phrase by means of an annexation of the word ***dhu*** (ذو) (owner of).

14 slots were reserved for Arabic adjectives as presented in table 6.1.b and it was decided to allocate 3 slots to visible quality, 3 slots to internal trait, 3 slots to an action, 3 slots to ***Nasab***, 1 slot to the word ***dhu*** (owner of) and 1 slot to be allocated for an adjective in the comparative form to represent the propriety of degree.

As described in the Arabic verb decomposition, a constraint was appended that one of the adjectives chosen must be part of a homograph adjective-verb pair. In order to promote the antonymy property (oppositeness of meaning), a constraint was appended that one of the adjectives selected should be opposite in meaning of one of the other adjectives in the frame.

Decomposition of the Arabic Adverbs

Rydin (2005) reported that "In Arabic, few words are adverbs in and of themselves and most words that function as Arabic adverbs are nouns or adjectives in the accusative case". (Rydin, 2005) decomposed the Arabic adverbs based on their semantic function into 7 classes which included adverbs of degree, manner, place, time, adverbial accusative of cause or reason, adverbial accusative of specification and compound time adverbials such as عندئذ "at that moment".

9 slots were reserved for Arabic adverbs as presented in table 6.1.b and it was decided to allocate 1 slot to each of the 7 classes, 1 slot for an adverb in the comparative form and 1 slot selected randomly. As discussed under the Arabic noun decomposition, a constraint was appended that one of the adverbs chosen must be part of a homograph adverb-noun pair.

6.2.1.2 Population of the Slots in the Frame with Arabic Words

The result of the decomposition process described in step 1 (section 6.2.1.1) is a tree structure which ranges from general Arabic words at the top level to specific

categories such as (relative, birds, and insects) at the lower level (leaves). These categories were used to derive the slots in the sampling frame. In accordance with the same process of filling the slots in the frame used in the AVSS-70 benchmark dataset (chapter 5 section 5.3.1.2); the list of Arabic word frequency (AWC) was partitioned into a high frequency list and a low frequency list. The high frequency list was created by selecting the most frequent 2000 words in the AWC list and the low frequency list contained the residue of the AWC list. The words in each list were randomised to avoid any bias. Each list was then separated into four sections (nouns, verbs, adjectives and adverbs) and each section subsequently searched for appropriate words to fill the slots based on their criteria specified through the process of the Arabic words decomposition.

As highlighted in chapter 5, an important issue in language representation is that of word frequency i.e. high frequency words should have a higher probability of appearing in the sample frame. For valid words representation, the decision was made to use the 80/20 rule used by (O'Shea et al., 2013) whereby 80% of the slots in the frame will be filled by random selection process with words from the high frequency list whilst 20% will come from the low frequency list. Table 6.5 illustrates the number of words that will be selected from high frequency list and from the low frequency list for each content words class.

Table 6.5 Frequency breakdown for Arabic content words classes

Content words classes	Words in class	Frequency breakdown
Nouns	30	24 High
		6 Low
Verbs	15	12 High
		3 Low
Adjectives	14	11 High
		3 Low
Adverbs	9	7 High
		2 Low

The result of this process is a set of 68 Arabic stimulus words which is presented in table 6.6 (LF means the word selected from low frequency list).

Table 6.6 The set of 68 Arabic stimulus words.

No	Class	Word	الكلمـة	Additional Criteria/Comments
1	Noun Abstract Human	Civilization	حضـارة	
2	Noun Abstract inanimate fact	Issue	قضـية	
3	Noun Abstract inanimate place	Capital-city	عاصـمة	
4	Noun Abstract inanimate action	Education	تعلـيم	
5	Noun Abstract inanimate time	Childhood	طفولـه	
6	Noun Abstract inanimate mental	Ability	قـدرة	
7	Noun Abstract inanimate emotion	Pride	فخـر	LF
8	Noun Abstract inanimate action	addition	إضافة	Selected randomly \ **Homograph Noun – Adverb pair with 60**
9	Noun Concrete Human Relative	Father	والـد	
10	Noun Concrete Human Military title	Officer	ضـابط	
11	Noun Concrete Human An occupation	Doctor	طـبيب	
12	Noun Concrete Human wise person	Messenger	رسـول	
13	Noun Concrete Human part of human body	Head	رأس	
14	Noun Concrete Human type of male's life stages	Lad	فتـى	
15	Noun Concrete Non-Human Animate Four footed animals	Lion	أسـد	
16	Noun Concrete Non-Human Animate Birds	Hawk	صقـر	LF
17	Noun Concrete Non-Human Animate Insect	Spider	عنكبـوت	LF
18	Noun Concrete Non-Human Animate Fish	Whale	حـوت	LF
19	Noun Concrete Non-Human Animate Snake	Viper	أفعـى	LF
20	Noun Concrete Non-Human Animate Diseases	Cancer	سرطـان	
21	Noun Concrete Non-Human Inanimate Living Tree	Date-Palm	نخـل	LF
22	Noun Concrete Non-Human Inanimate Living Flower	Rose	وردة	LF
23	Noun Concrete Non-Human Inanimate Non-Living Metal	Gold	ذهـب	**Homograph noun-verb pair With 33**

24	Noun Concrete Non-Human Inanimate Non-Living Reading material	Book	كتـاب	
25	Noun Concrete Non-Human Inanimate Non-Living building for religious services	Masjid	مسجـد	
26	Noun Concrete Non-Human Inanimate Non-Living Weapon	Rifle , Gun	بندقيـة	
27	Noun Concrete Non-Human Inanimate Non-Living Weather phenomenon	Earth quake	زلزال	
28	Noun Concrete Non-Human Inanimate Non-Living Transportation vehicle	Aircraft	طائـرة	
29	Noun Concrete Non-Human Inanimate Non-Living Non-alcoholic beverage	Milk	لبــن	
30	Noun Concrete Non-Human Inanimate Non-Living Part of day	Dawn	فجــر	
31	Verb Action Basic	Take on	إتخـذ	AVN verb class IisoTaEomala-1
32	Verb Action Experiential	Reveal	كشـف	AVN class Oazohaea-1(also Bay~ana-1)
33	Verb Action Locative	Go	ذهــب	AVN class Haraba-1 (also Iinotahay-1, Other-cos-1, Iixotafay-1) \ **Homograph noun-verb pair with 23**
34	Verb Action Benefactive	Reward	كافـأ	Jah~aza-1.2 \ **Paired with Honour 35**
35	Verb Action	Honour	أكــرم	**LF** \ AVN class Jah~aza-1.2 \ **Paired with Reward 34 \ Homograph verb-adjective pair with 59**
36	Verb State Basic	Be-issued	صــدر	AVN class Nataja-1 (also Hasala-1, Zahara-1, Oarosala-1)
37	Verb State Locative	Be-connected	إرتبط	AVN class Iilotasaqa-1
38	Verb State Benefactive	Have\ Own	مـلك	AVN class Malak-1and OaEotay-1
39	Verb State Experiential	Be-glad	فرح	AVN class Tasal~ay-1.3\ **Paired with Rejoice**
40	Verb State	rejoice	إبتهـج	**LF** \ AVN class Tasal~ay-1.3 \ **Paired with Be-glad 39**
41	Verb Process Experiential	Excite	أثــار	AVN class Sab~aba-1
42	Verb Process Basic	Melt	ذاب	AVN class Iimotazaja-1.2
43	Verb Process Locative	remain	بقــي	AVN class Wujida-1 and Oaqaama-1
44	Verb Process Benefactive	Gain	ربــح	AVN class Ii$otray-1.1.1 \ **source of 3th level class \ Paired with acquire 45**

45	Verb Process	Acquire	تكسب	**LF** \ AVN class Ii$otray-1.1.1 \ **Paired with gain 44**
46	Adjective Visible Quality	Blue	أزرق	
47	Adjective Visible Quality	Voluminous	ضخـم	
48	Adjective Visible Quality	New	جديـد	
49	Adjective An action	Specialized	متخصص	
50	Adjective An action	Emigrating	مهاجر	
51	Adjective An action	Standing	قائـم	LF
52	Adjective Internal Trait	Generous	كريـم	Source for comparative (with 59)
53	Adjective Internal Trait	Intelligent	ذكـي	
54	Adjective Internal Trait	Envious	حسـود	LF
55	Adjective Nasba	International	دولـي	Source for antonym, Local (with 57)
56	Adjective Nasba	Spatial, Space	فضائـي	
57	Adjective Nasba	Local	محلـي	Antonym of International (with 55)
58	Adjective descriptive phrase through an annexation of the word ذو (owner of)	Rich, wealthy	ذو مال	
59	Adjective Comparative	More generous	أكـرم	**LF** \ comparative of generous (with 52) \ **Homograph verb-adjective pair with 35**
60	Adverb of manner \ Noun in accusative	In addition	إضافـةً	**Homograph noun-adverb pair with 8**
61	Adverb of degree \ Basic adverb	Only	فقط	
62	Adverb of time \ Noun in accusative	Morning	غـدوةً	LF \ **same class with noun (Dawn) 30**
63	Adverb of place	near	قـرب	Source of comparative (with 68)
64	Adverbial accusative of cause or reason Noun in accusative	In preparation for	تمهيـداً لـ	
65	Adverbial accusative of specification\ Adjective in accusative	Economically	إقتصـادياً	
66	Compound time adverbial	At that time	عندئذ	
67	Adverb of degree \ adjective in accusative	completely	تمامـاً	Randomly selected
68	Adverb Comparative	nearer	اقرب	**LF** \ comparative of near (with 63)

6.2.2 Production of the Arabic Short Text Pairs

The second stage of the creation of the ASTSS-68 dataset methodology was the production of the set of 68 Arabic short text pairs. The methodology of the production of this set consisted of:

1. Creation of a database of 1088 Arabic short texts.
2. Selecting candidate pairs of short texts of high and medium similarity from the created database by three judges.
3. Conduct an experiment to select the final set of 68 short text pairs from the set of the candidate pairs selected by judges, plus a randomly selected a set of low similarity short text pairs from the database.

6.2.2.1 Creation of the Arabic Short Text Database Experiment

Participants

The original aim of this experiment was to create a database of 1088 Arabic short texts to be used later for selecting the set of 68 Arabic short text pairs. In order to balance the efforts of the Arabic participants against the number of generated short texts, this experiment used a sample of 32 Native Arabic speakers. Each participant was asked to write two short texts derived from 17 Arabic stimulus words. This sample size would produce ((17x2) x32 = 1088) Arabic short texts offering scope to find appropriate similarity combinations. The experiment required participants with a capacity for creative writing therefore the decision was made to use a sample of participants within the following disciplines: Arabic linguistics, journalism, writing, Arabic teaching, religious sciences and religious teaching.

The participants were from 8 Arabic countries which included: Iraq, Saudi Arabia, Egypt, Jordan, Lebanon, Syria, Morocco, and Palestine. Each of the 32 participants received a questionnaire (described later) with instructions to be followed to generate the short texts. The participants who lived outside the UK received the questionnaire by email whilst those inside the UK received it by post.

Materials

In order to assemble the database of 1088 Arabic short texts, the set of 68 Arabic stimulus words was partitioned into 4 blocks of 17 stimulus words which were A_1, A_2, B_1 and B_2. The process of collecting this database included writing two short texts for each of the 17 stimulus words in a specific block. The purpose of partitioning the set of stimulus words into 4 blocks was to distribute the workload and to avoid spurious semantic overlap. For example, the comparative adjective اكرم "more generous" appeared in block A_1 whilst the adjective كريم "generous" appeared in block B_1. However, the verb اكرم "honour" (homograph with the comparative adjective اكرم "more generous") appeared in B_2 so that no pair of short texts selected from two different blocks could have been written by the same person (or different people experiencing the same semantic context). The full block structure for this experiment is presented in table 6.7. Where, LF means low frequency, V- verb, N- noun, Adj- adjective, Adv- adverb, A- abstract, R- randomly selected, Comp- comparative and Ant- antonymy.

Some stimulus words were selected from the same or related classes to promote high and medium similarity such as the verbs ابتهج "rejoice" and فرح "be glad". However, the issue of obtaining a large number of low similarity short texts was still possible. An additional constraint was added with some stimulus words to solve this problem using the thematic similarity (Klein and Murphy, 2002) which is alternative approach to semantic similarity. (Mirman & Graziano, 2012) reported that "concepts whose similarity is based on frequent co-occurrence in situations or events are thematically related, such as dogs and leashes do not share features and are not members of the same category, but both are frequently involved in the taking-the-dog-for-a-walk event or situation".

It was decided to use the thematic similarity with some Arabic stimulus words of each content word class whereby participants were requested to write two short texts using the stimulus word within a specific theme. The thematic similarity was used based on the assumption that two short texts produced using the same word and the same theme were probable to have a high level of similarity whilst the short texts

using either the same stimulus word or the same theme were probable to have a medium similarity.

Table 6.7 Blocked design to distribute materials to participants.

A1	A2	B1	B2
Noun 7	Noun 8	Noun 8	Noun 7
Ability A قدرة	Issue A قضية	Education A تعليم	**Addition A (N-Adv) R إضافة**
Father والد	Pride A LF فخر	Childhood A طفولة	Messenger رسول
Lad فتى	Officer ضابط	Head رأس	Spider LF عنكبوت
Hawk LF صقر	Lion اسد	Viper LF أفعى	Whale LF حوت
Gold (N-V) ذهب	Date Palm LF نخل	Rose LF وردة	Book كتاب
Milk لبن	Rifle بندقية	Masjid مسجد	Earth quake زلزال
	Aircraft طائرة	**Dawn (same noun class with adverb morning) فجر**	
Civilization A حضارة	Cancer سرطان	Doctor طبيب	Capital–city A عاصمة
***Travel and tourism* السفر والسياحة**	***Health and happiness* الصحة والسعادة**	***Health and happiness* الصحة والسعادة**	***Travel and tourism* السفر والسياحة**
Verb 4	Verb 4	Verb 3	Verb 4
Take on إتخذ	Reveal كشف	**Go (N-V) ذهب**	**Honour LF J1.2 (V-Adj) أكرم**
Be issued صدر	Be connected إرتبط	Remain بقي	Have\Own ملك
Acquire LF I_1.1.1 تكسب	Melt ذاب		Excite أثار
Be glad T_1.3 فرح	**Reward J1.2 كافأ**	**Rejoice LF T1.3 إبتهج**	**Gain I_1.1.1 ربح**
***The Muslim festivals* اعياد المسلمين**	***Sport and leisure* الرياضة والترفيه**	***The Muslim festivals* اعياد المسلمين**	***Sport and leisure* الرياضة والترفيه**
Adjective 4	Adjective 3	Adjective 4	Adjective 3
More generous (V-Adj) LF أكرم	Blue أزرق	Voluminous ضخم	New جديد
Emigrating مهاجر	Specialized متخصص	Standing LF قائم	Envious LF حسود
Rich \ wealthy ذو مال		**Generous (comp) كريم**	
Spatial \ space فضائي	**International (Ant) دولي**	Intelligent ذكي	**Local (Ant) محلي**
***News and media* الاخبار و وسائل الاعلام**	***News and media* الاخبار و وسائل الاعلام**	***News and media* الاخبار و وسائل الاعلام**	***News and media* الاخبار و وسائل الاعلام**
Adverb 2	Adverb 2	Adverb 2	Adverb 3
In addition (N-Adv) إضافةً	Only فقط	Near قرب	In preparation for تمهيداً لـ
Economically اقتصادياً	**Morning غدوةً (same noun class) LF**	At that time\moment عندئذ	Completely **R** تماماً
			Nearer (comp) اقرب

Themes were used with some stimulus words of each content word class (nouns, verbs and adjectives) apart from the adverbs. The majority of stimulus adverbs comprised nouns or adjectives in the accusative case which appear in a short text within a specific context. It was likewise decided not to add themes to adverbs because additional constraints could make production of short texts infeasible or artificial.

The use of thematic similarity needed a suitable source of themes, which were chosen from language instruction texts for non-native Arabic speakers (Smart, 1992, Wightwick and Gaafar, 2007 and 2009) which concentrate on talk about useful everyday activities. An example of these themes is اعياد المسلمين "the Muslim

festivals" which was used with the verbs ابتهج "rejoice" and فرح "be glad" as shown in table 6.7. A full list of themes is presented in appendix 6.

In this research, themes were chosen based on general occurrence and possibility of being useful with the Arabic stimulus words. Themes chosen for use in this experiment were the Muslim festivals, travel and tourism, health and happiness, sport and leisure, and news and media. These themes were used to encourage the production of high similarity short text pairs through use of the same stimulus word. Also to encourage the production of medium similarity short text pairs by applying the theme to two stimulus words selected from the same or related class. As shown in table 6.7, the health and happiness theme was applied to the stimulus words طبيب "doctor" and سرطان "cancer".

Instructions and Procedure

32 participants divided into four groups of eight Native Arabic speakers took part and generated 1088 short texts. The participants in each group received questionnaires with instructions as to how to generate short texts using 17 stimulus words allocated to a specific block (A_1, A_2, B_1 or B_2). The words in each questionnaire were categorized based on their class (nouns, verbs, adjectives and adverbs). Each class commenced with an instruction page containing a definition of the word class with examples. Each stimulus word was accompanied by instruction to write two short texts and printed on a separate sheet with boxes for responses.

Each questionnaire had three themes which were applied to the final noun, final verb and final adjective presented to the Arabic participants. Extra sheets were added to explain the task before the final noun, final verb and final adjective. The participants were asked to write two short texts of between 10 to 20 words in length in clear handwriting using the stimulus word (and on the general topic, if a specific theme applied)). Appendices 7.1 and 7.2 contain a sample extracted from the questionnaire which include instruction sheet and sheets to explain the task and to write the two short texts with and without theme, respectively.

Some information (appendix 7.1) was added to assist the participants in how to treat the homograph pairs and the polysemous words together with notes to encourage the participants to use all types of the dialogue acts such as question, instruction, statement, etc. An additional sheet was added to clarify the difference between instructions, expressions, statements, commitments or declaration. Appendix 7.3 contains an example of this sheet.

The result of this experiment was a database of 1088 Arabic short texts written by 32 Arabic participants.

6.2.2.2 Selection of the Set of 68 Short Text Pairs

The created database of 1088 Arabic short texts was used to select a set of 68 Arabic short text pairs. In order to accomplish this, 130 queries were presented to extract groups of short texts from the created database. These queries were generated based on criteria used for allocating the stimulus words in order to provide different degrees of similarity and also to ensure that each stimulus word was likely to appear in the final dataset at least once. These included generating queries which would return all short texts produced for a particular Arabic stimulus word, all short texts for each pair of Arabic stimulus words (e.g. father and lad which have the common features Noun:Concrete:Human) and all short texts produced by a particular theme such as ***Travel and tourism***. Queries were also produced to select short texts for pairing set between the stimulus words in blocks A_1, A_2, B_1 and B_2. If a pair contains two short texts from the same stimulus word, extra checking is required to make sure they come from different authors.

The process of selecting the final set of Arabic short text pairs consisted of two steps.

6.2.2.2.1 Selection of the candidate short text pairs by judges

This step included selecting candidate pairs of short texts of high and medium similarity from the created database by a committee of three judges. Two of the judges were Arabic linguistics and the third was an Arabic speaking expert in

semantic similarity. Each judge was provided with the printed queries. The queries were examined by each judge who was required to nominate two sets of short text pairs in isolation which included high and medium similarity. On account of the difficulty in convening a meeting of all judges to agree on the sets of the short text pairs selected, it was decided to select the pairs of short texts that had been nominated by all 3 judges as high similarity and medium similarity pairs. The pairs which were nominated by two judges were printed on a separate sheet and sent again to the third for the purpose of reaching a consensus. Based on the judgements, a set of 65 pairs of high and medium similarity short texts was identified. The set consisted of 29 candidate pairs of high similarity short texts and 36 candidate pairs of medium similarity short texts.

6.2.2.2.2 Selection of the final short text pairs experiment

Because the judges had difficulty in reducing the medium similarity candidate pairs to a coherent set and also selection of candidate high and medium similarity pairs by human judges in the past has not always been effective (O'Shea et al., 2013), it was decided to use a new sample of 10 participants in an experiment to select a sample of high and medium similarity pairs with greater confidence, before running the rating experiment. Another very low similarity short text pair was added to the set of 65 short text pairs selected by judges (making 66 short text pairs), to ensure that the selectors saw the full similarity range and did not bias their selection the final set.

The card sorting technique with semantic anchors which was identified in the creation of the ANSS-70 dataset as most suitable was used in this experiment to collect human ratings for 66 pairs of short texts.

The sample of 10 native Arabic speakers used in this experiment was from 3 Arabic countries (Iraq, Saudi Arabia and Libya). Each of the 66 short text pairs was printed on a separate card and these cards were presented to participants for rating how similar in meaning were the short texts on each card. Each of 10 participants was supplied with an envelope containing 66 cards and three sheets which included: an instruction sheet to collect human ratings, a sheet to record the similarity ratings and

a sheet for the personal information. The order of 66 cards was randomized before presentation to reduce the ordering effects.

The same procedure was followed as in the ANSS-70 dataset to collect human ratings. The participants were asked to sort the cards into four groups in accordance with the similarity of meaning. The High Similarity of Meaning (HSM) group contained short text pairs between strongly related and identical in meaning. The high Medium Similarity of Meaning (MSM) group contained short text pairs very much alike in meaning, whilst the low MSM group contained pairs which were vaguely similar in meaning and the Low Similarity of Meaning (LSM) contained pairs unrelated in meaning. After sorting the cards, the participants were asked to check them carefully and then rank each short text pair using a point on a rating scale described by the semantic anchors which ran from 0.0 (unrelated in meaning) to 4.0 (identical in meaning). The instruction sheet also included some notes which enabled participants assigning an accurate degree of similarity by means of use of the first decimal place and to avoid using values lower than 0.0 or greater than 4.0 to rate the short text pairs.

Table 6.8 illustrates the outcome of this experiment. The final set of 68 short text pairs was selected based on the experimental results plus randomly selected low similarity short text pairs from the database as follows:

Table 6.8 the distribution of similarity ratings in the set of 66 short text pairs.

Similarity Range	**Number of Short Text Pairs**
0.00 – 0.99	4
1.00 – 1.99	14
2.00 – 2.99	19
3.00 – 4.00	29

1. It was decided to randomly select the 22 of 29 short texts that were rated high by all 10 participants to represent the high similarity short text pairs in the final set.

2. To obtain a good similarity range representation in the final set, 23 short texts rated medium by participants within the range (1.00 – 2.50) were selected to represent the medium similarity pairs in the final set. The reason for this choice was that some of the pairs in the broad medium band (1.00 - 2.99) with ratings of over 2.50 were rated high by more than half the participants. These pairs may still get a high similarity rating when the participants are increased to 60 in the final stage of the creation of this dataset.
3. Finally, 23 pairs were chosen as a combination of 4 short text pairs rated low by participants plus 19 pairs selected randomly from the database to represent the low similarity pairs in the final set. These were scrutinised to check that no obvious medium or high similarity combinations had occurred by chance.

6.2.3 Collecting the Similarity Ratings for 68 Short Text Pairs

The card sorting technique with semantic anchors used in the experiment of the selection of the final set of Arabic short text pairs was employed in this section to collect human similarity ratings for the produced set of 68 Arabic short text pairs. The process of collecting human similarity ratings involved two steps:

6.2.3.1 Pilot Study

The aim of the pilot study was to investigate whether the 68 short text pairs arrived at by the process in 6.2.2.2.2 had a good representation of the similarity range before committing to a large-scale ratings experiment. A new sample of 8 native Arabic speakers from four Arabic countries (Iraq, Saudi Arabia, Libya and Jordan) was used in this experiment.

Each of the 68 short texts was printed on a separate card. Each of 8 participants was supplied with an envelope containing 68 cards and three sheets (as in the experiment for the selection of the final set of Arabic short text pairs) which included: an instruction sheet to collect human ratings, a sheet to record the similarity ratings and a sheet for the personal information. The order of the 68 cards was randomized before presentation to reduce the ordering effects. The participants were asked to rate

the 68 short text pairs on how similar they were in meaning using card sorting and semantic anchors.

Table 6.9 illustrates the result of this experiment which indicates that the set of 68 short text pairs achieved a good balance in the number of short text pairs of each similarity range apart from one pair of high similarity short texts which was rated as medium by participants. Consequently, this pair was replaced with another one that rated high by both judges and the participants in the experiment of the selection of the final set of short text pairs.

Table 6.9 the distribution of similarity ratings in ASTSS-68 dataset pilot study.

	Before Pilot Study	**After Pilot Study**
Similarity Range	**Number of Short text Pairs**	**Number of Short text Pairs**
Low similarity	23	23
Medium similarity	23	24
High similarity	22	21

6.2.3.2 Conduct of the Final Ratings Collection Trial

The decision was made to include the ratings collected from the pilot trial in the final study experiment. The 8 participants were asked again to rate each of the new pairs of short texts which were added after the outcome of the pilot study was reviewed. This experiment used a new sample of 62 participants including 8 participants from the pilot study. This sample was chosen based on experience with the previous experiment of ANSS-70 benchmark dataset which indicated that the sample of 60 participants was adequate for the obtainment of good quality ratings. The sample was chosen on the basis of its being a general population with equal balance between students and non-students.

1. All were Arabic native speakers from 7 Arabic countries which included: Saudi Arabia (15), Iraq (14), Syria (10), Libya (9), Palestine (6) Egypt (4), and Jordan (4).

2. The participants' academic backgrounds consisted of 38 Science/Engineering vs. 24 Art/Humanities. In case of educational level, the balance was obtained and the overall breakdown qualifications were illustrated in table 6.10.

Table 6.10 participants' educational background

Student	**Non-student (highest qualification)**
14 undergraduate	15 Bachelors
7 Masters	6 Masters
10 PhD	5 PhD
None	5 secondary school

3. Equally balance was achieved between female and male. The gender balance achieved for non-student was (16 male and 15 female) whilst for student (15 male and 16 female).

4. In case of age, Table 6.11 shows the age distributions of a selected sample.

Table 6.11 Age distributions for the Arabic population sample.

Age range	**Participants**	
18-22	14	Student 14
		Non-student 0
23-29	9	Student 2
		Non-student 7
30-39	27	Student 10
		Non-student 17
40-49	10	Student 5
		Non-student 5
50-59	2	Student 0
		Non-student 2

The participants followed the same procedure as had been undertaken to collect human ratings in the pilot study. They were asked to rate 68 short text pairs using the card sorting and semantic anchors.

6.2.3.2.1 Experimental Results and Discussion

Table 6.12 represents the results of this experiment which contains the set of 68 Arabic short text pairs with human ratings of similarity. The human similarity ratings collected in this experiment were calculated as the mean of the judgements provided by the 62 Arabic native speakers for each pair of short texts. The second and last columns represent the set of Arabic short text pairs in Arabic with approximate translation to English. The third column contains the mean of similarity rating collected from 62 Arabic native speakers whilst the fourth column represents the Standard Deviation (SD) of each short text pair which demonstrates an inevitable degree of noise in human ratings.

The approximate translations of the Arabic short texts have not been made to good colloquial English – rather they are literal translations which help the English reader to map the processes taking place onto the original Arabic texts.

Table 6.12 Arabic Short Text Benchmark Dataset (ASTSS-68)

ST	Short Text Pairs	Human Ratings	SD	ازواج الجمـــل العربيـــة
1	Muslims are happily celebrated with Eid Al-Ftir because it comes after a long month of fasting and self-strive.	3.80	0.43	يفرح المسلمون بعيد الفطر لانه ياتي بعد شهر طويل من الصيام وجهاد النفس.
	Fasting people rejoice the blessed Eid Al-Ftir which is a reward for them after a month of fasting.			يبتهج الصائمون بعيد الفطر المبارك الذي يعد مكافأة لهم بعد مشقة شهر من الصيام.
2	O son, I advise you not to earn your livelihood from illegal work because it conceals the blessing from it.	3.38	0.70	انصحك يابني ان لا تتكسب قوتك من الحرام لان المال السحت فيه ممحقة للبركة.
	Allah blesses the man who makes living from legal work, as he prohibited the illegal money.			يبارك الله في الرجل الذي يتكسب ماله من الحلال لانه حرم المال السحت.
3	South Cairo court ruled in the case of the killing of the demonstrators last Thursday.	2.12	0.81	قامت محكمة جنوب القاهرة بالحكم في قضية قتل المتظاهرين الخميس الماضي.
	The recent report of the fact-finding committee revealed the involvement of some of the remnants of the former regime in the killing of the demonstrators.			كشف تقرير لجنة تقصي الحقائق الاخير عن تورط بعض رموز النظام السابق في قتل المتظاهرين.
4	A professional football player earns a lot of money from the club he plays for and from a competition and thus he enjoys living luxurious life.	1.66	0.84	يتكسب لاعب كرة القدم المحترف الكثير من المال من النادي الذي يلعب له والبطولات التي يشارك فيها فينعم بحياة مترفة.
	After he won a large sum of money, the tennis player travelled with his family on a trip for the purpose of entertainment and recreation.			بعد ان ربح لاعب التنس البطولة مع مبلغ كبير من المال سافر في رحلة مع عائلته للترفيه والاستجمام.
5	I work in the university teaching in addition to my work in the linguistic assessment of books and publication in literary works.	0.01	0.06	اعمل في التدريس الجامعي اضافة الى عملي في التقويم اللغوي للكتب والمنشورات في الاعمال الادبية.
	I got cold which resulted in coughing and my mother advised me to add a spoon of honey to the lemon juice, which will help me a lot in getting better.			اصبت بالبرد ونتج عنه السعال فنصحتني والدتي بأضافة ملعقة عسل الى عصير الليمون يسارع كثيرا في شفائي.

6	Take the friend a faithful brother, honest with you and will help you in the time of adversity.	**0.08**	**0.27**	**اتخذ الصديق اخا وفيا لك صادقا معك يعينك في وقت المحن والشدائد.**
	Do not give any judgement when you are in a state of anger because anger is a silent demon.			**لا تصدر حكما وانت في حالة غضب لأن الغضب شيطان أخرس.**
7	The Iraqi team was about to win the Arab Gulf Football Championship cup except for sudden loss to the United Arab Emirates team	**3.43**	**0.69**	**كاد الفريق العراقي ان يربح كأس بطولة الخليج العربي لكرة القدم لو لا خسارته المفاجئة امام الفريق الاماراتي .**
	The Emirati team won the final match of the last Gulf Cup, which took place in Bahrain and deservedly won the championship cup.			**ربح المنتخب الاماراتي المباراة النهائية لبطولة الخليج الاخيرة التي جرت في البحرين وكسب كأس البطولة بجدارة.**
8	Milk is wholesome food and it is necessary for children and adults to have it as it builds bones because it is rich with calcium.	**3.61**	**0.61**	**اللبن غذاء نافع ومن الضروري ان يتناوله الاطفال والكبار لبناء عظام الاجسام فهو غني بمعدن الكالسيوم.**
	A lot of people eat yogurt for the purpose of obtaining calcium to strengthen and protect their bones.			**يتناول الكثير من الأشخاص اللبن طمعا في الحصول على الكالسيوم لتقوية العظام وحمايتها.**
9	The black widow spider is famous for its poison which affects the nerves and it is available all over the world.	**3.81**	**0.42**	**عنكبوت الأرملة السوداء أحد العناكب المشهورة بسمها المؤثر على الأعصاب و يتواجد في جميع دول العالم.**
	The black widow is a kind of large-sized spider with a deadly poison.			**تعد الارملة السوداء نوعا من انواع العناكب الكبيرة الحجم ذات السم القاتل.**
10	Sky today is blue and clear unlike yesterday as it was cloudy.	**2.4**	**0.78**	**السماء اليوم زرقاء اللون صافية على عكس الأمس فقد كانت ملبدة بالغيوم.**
	How beautiful is it that the sky is blue, the sun is shining, and the sea still with little white clouds here and there.			**ما أجمل أن تكون السماء زرقاء و الشمس مشرقة والبحر ساكنا مع القليل من الغيوم البيضاء هنا وهناك.**
11	I will meet you early in the morning between dawn and sunrise.	**1.31**	**0.79**	**سألقاك في أثناء الغدوة أي ما بين الفجر وطلوع الشمس أول النهار.**
	How wonderful for you to wake up early before dawn and the spread of light as that increases your energy			**ما أجمل أن تستيقظ مبكراً قبل طلوع الفجر وانتشار الضياء يزيدك ذلك نشاطا وحيوية طول اليوم.**

	throughout the day.			
12	Iraq has been named the land of blackening for the intensity of the greenery and fertility in addition to the large number of palm trees in its land.	**0.05**	**0.21**	**اطلق على العراق ارض السواد لاشتداد الخضرة والخصب فيه و لكثرة النخل في ارضه.**
	I offer you this rose in recognition of my gratitude to the great what you have done.			**أقدم لك هذه الوردة عرفاناً بجميل ما صنعت وعظيم ما أسديت.**
13	Iraq witnessed economical and commercial growth after the discovery of oil in large quantities in its land.	**0.23**	**0.42**	**شهد العراق نموا اقتصاديا وتجاريا بعد اكتشاف النفط بكميات كبيرة في اراضيه.**
	The state has set up huge dams to store rain to be utilized in various fields.			**اقامت الدولة سدودا ضخمة لخزن الامطار وذلك للاستفادة منها في شتى المجالات.**
14	Beware of the using the hunting rifle in front of children because they will perceive it as a toy and that may put an end to their life.	**3.83**	**0.41**	**حذاري من استعمال بندقية الصيد امام الاطفال فانهم سيتصورونها لعبة وقد تنهي حياتهم.**
	Do not leave a rifle in a place that children can reach, it is very dangerous.			**لا تترك البندقية في مكان يتمكن الأطفال من الوصول إليه فهي خطرة جدا.**
15	Cancer is one of serious diseases of the age that still represents a challenge for doctors and patients.	**3.33**	**0.61**	**مرض السرطان احد امراض العصر الخطيرة التي مازالت تُمثِّل تحدياً للأطباء و المرضى.**
	Cancer is considered as one of the most serious diseases that affect the health and happiness of the individual.			**يعتبر مرض السرطان من أكثر ألامراض التي تؤثر على صحة وسعادة الفرد.**
16	Muslim strives hard to pray the dawn prayer at the time and in the Masjid because it grants him a great reward.	**3.54**	**0.83**	**يجتهد المسلم كي يصلي الفجر في وقتها وفي المسجد لما في ذلك من اجر عظيم.**
	The dawn prayer is one of the important prayers for Muslims and it should be done on time.			**تعد صلاة الفجر من الصلوات المهمة عند المسلمين ويجب ان تصلى في وقتها.**
17	I feel proud of my son’s success in his study and distinctiveness over his colleagues.	**2.35**	**0.95**	**اشعر بالفخر والاعتزاز بنجاح ابني في دراسته وتميزه على زملائه.**
	Would you feel happy and proud if you knew that one of your students became the ruler of the country?			**اما تحس بسعادة وفخر اذا علمت ان احد طلابك اصبح حاكما للبلاد ؟**

18	Originally praying is to be done by a Muslim while standing and can be done while he is sitting for those people who have legitimate excuses	**1.44**	**0.88**	**الاصل في الصلاة ان يؤديها المسلم قائما ويصح الجلوس فيها لذوي الاعذار الشرعيه.**
	Neighbor of the Masjid must pray in the Masjid unless there is a legitimate excuse for it			**جار المسجد لاتقبل منه صلاته الا في المسجد مالم يكن هناك عذر شرعي لذلك.**
19	Please, the games in this stadium are intended only for children under the age six.	**0.07**	**0.25**	**من فضلكم الالعاب في هذا الملعب مخصصة للأطفال دون سن السادسة فقط.**
	Do not use a mobile phone while driving a car because you may be exposed to a serious accident.			**لا تستخدم الهاتف النقال اثناء قيادة السيارة لانك قد تتعرض عندئذ لحادث مؤسف.**
20	Do not remain exposed to the oblique sunlight for a long period because it leads to skin cancer.	**0.05**	**0.28**	**لا تبق معرضا الى ضؤ الشمس المائل فترة طويلة لانها تؤدي للاصابة بامراض سرطان الجلد.**
	Despite the passage of thousands of years, there are still traces of ancient civilizations based on our land up to this day.			**برغم مرور الاف الاعوام ما زالت آثار الحضارات القديمة قائمة الى يومنا هذا.**
21	The last messenger sent by Allah to all mankind, told the message to the fullest and did all what Allah had commanded him.	**2.98**	**1.06**	**إن آخر رسول أرسله الله إلى البشر اجمع أبلغ الرسالة على أكمل وجه وأتم جميع ما أمره الله به .**
	Almighty Allah sent Mohammad, peace be upon of him, a messenger to all people and worlds to take them out of darkness and into the light.			**ارسل الله سبحانه محمدا صلى الله عليه وسلم رسولا الى الناس كافة والعالمين ليخرجهم من الظلمات الى النور.**
22	The media in each country is concerned with local news as it is concerned with international news.	**3.23**	**0.77**	**وسائل الاعلام في كل دولة تهتم بالخبر المحلي مثلما تهتم بالاخبار الدولية.**
	Local media is always concerned with the internal news more than the world news.			**وسائل الاعلام المحلية تعنى دائما بالاخبار الداخلية للبلد اكثر من الاخبار العالمية.**
23	Local media quoted a story that the traffic police have organized the process of vehicle traffic in the streets of the capital which is witnessing a major traffic jam.	**2.11**	**0.88**	**نقلت وسائل الاعلام المحلية خبرا مفاده ان شرطة المرور قاموا بتنظيم عملية سير المركبات في شوارع العاصمة التي تشهد اختناقا مروريا كبيرا.**

	A friend told me about a great traffic accident in the capital at the moment and when he felt my surprise, he said that I subscribe in the breaking news service via smart phones.			اخبرني صديقي عن وقوع حادث مروري كبير في العاصمة في هذه اللحظة وعندما شعر باستغرابي قال انا اشترك بخدمة الاخبار العاجلة للهواتف الذكية.
24	There must be donated to support and assist victims of the earthquake that hit Turkey.	2.50	0.79	يجب المساهمة بالتبرع ماديا لدعم ومساعدة ضحايا الزلزال الذي ضرب تركيا.
	Islamic relief organizations decided to donate a significant amount to help countries suffering from the famine.			قررت منظمات الاغاثة الاسلامية التبرع بمبالغ ضخمة لمساعدة الدول التي تعاني من المجاعة.
25	India pledged to send a Cobra snake to the zoo in Al-Zewra park in a glass basin.	1.73	0.82	تعهدت الهند بارسال افعى الكوبرا الى حديقة الحيوانات في متنزه الزوراء بحوض زجاجي.
	Large numbers of the dangerous venomous Cobra spread over in India, which its poison is considered as the most deadly one and can kill a person within few seconds.			تنتشر في الهند اعدادا كبيرة من افعى الكوبرا الخطرة التي يعد سمها من أنواع السموم القاتله التي تكفي لقتل الشخص في ثواني.
26	Raising dust provokes allergies in many people who suffer respiratory problems.	0	0	الغبار المتطاير يثير الحساسية لدى الكثير من الاشخاص الذين لديهم مشاكل في الجهاز التنفسي.
	Do not reveal your secrets to everyone and you become vulnerable to blame.			لا تكشف اسرارك لكل من هب ودب فتصبح عرضة للملامة.
27	Our company has the ability to manufacture quality home furniture and deliver it to customers in a short period of time.	0.03	0.18	لدى شركتنا القدرة على صناعة الاثاث المنزلي الجيد وتسليمه للزبائن خلال مدة قياسية.
	Make your lecture take two hours and then I will pay you an amount that you have never received.			اجعل محاضرتك تستغرق ساعتين عندئذ ادفع لك اجرا مجزيا لا عهد لك به.
28	The minister rewarded the players who got the gold medal in London Olympics.	3.34	0.74	اكرم الوزير اللاعبين اللذين حصلوا على الميدالية الذهبية في اولمبياد لندن.
	The Ministry of Youth and Sport decided to offer a reward for each player to win a medal in the next Olympics.			قررت وزارة الشباب والرياضة ان تكافأ كل لاعب يفوز بميدالية في الاولمبياد القادمة.
29	Most oriental women have large quantities of gold which they use for decoration and as a saving.	3.50	0.64	اغلب النساء الشرقيات يملكن كميات كبيرة من الذهب ويستخدمنه للزينة والتوفير.

	Gold is the best ornament for oriental women so they purchase it heavily.			يعد الذهب من افضل الحلي لدى النساء الشرقيات لذلك يُكثرن من شراءه.
30	If I take my children with me to the zoo, I will not let them put their hands in the lion's cage.	3.34	0.62	اذا اخذت ابنائي معي الى حديقة الحيوانات فلن اتركهم يضعون ايديهم في قفص فيه اسد.
	Baby, do not approach a lion's cage it will nibble your soft hand.			لا تقترب يا صغيري من قفص الاسد كي لا يقضم يدك الناعمة.
31	Do not pick a rose from public parks in your city so as not to deprive others of the enjoyment of its beauty.	2.23	0.75	لا تقطف وردة من الحدائق العامة في مدينتك لكي لا تحرم الاخرين من التمتع بجمالها.
	The concerned parties in the capital cultivated a thousand roses in the public parks.			قامت الجهات المسؤولة في العاصمة بزراعة الف وردة في الحدائق العامة.
32	I've done all my required work in addition to contributing to some charity works.	1.81	0.76	لقد انجزت جميع اعمالي المطلوبة مني اضافة الى المساهمة في بعض الاعمال الخيرية.
	The employee should perform his duties faithfully in addition to respecting the work schedule.			على الموظف ان يؤدي واجباته باخلاص اضافة الى احترام مواعيد الدوام.
33	You must be a messenger of good if you want to reconcile between the opposing parties.	0.19	0.42	يجب ان تكون رسول خير اذا اردت ان تصلح بين الاطراف المتخاصمة.
	Will the issue of Sheikh Ahmed be discussed this afternoon in the conference hall?			هل ستناقش قضية الشيخ أحمد اليوم بعد العصر في قاعة المؤتمرات ؟
34	You should consult a doctor specializing in the disease and he will give you the right cure by God's will.	0.03	0.18	عليك استشارة الطبيب المتخصص بالمرض فهو سيعطيك الدواء الشافي بإذن الله.
	I decided to sell my rifle after the issuance of the new law to prevent the posession of weapons.			قررت أن أبيع بندقيتي بعد إصدار القانون الجديد بمنع حيازة الأسلحة.
35	Hatim Al-Tai is the most generous person known by the Arabs and was mentioned in history books.	3.89	0.32	حاتم الطائي هو اكرم شخص عرفه العرب وذكرته كتب التاريخ.
	In the history of man, the Arab nation did not know a generous man more than Hatim Al-Tai			لم تعرف أمة العرب في تأريخها رجلا أكرم من حاتم الطائي.

36	The blue whale lives in the seas and oceans and feeds on small fish and plankton that enter his mouth with water.	**3.70**	**0.45**	**يعيش الحوت الازرق في البحار و المحيطات ويتغذى على الاسماك الصغيرة والعوالق البحرية التي تدخل فمه مع المياه .**
	The blue whale is the largest animal on earth and has no teeth but strongly rushes into the water to feed on the sea floating livings.			**يعد الحوت الأزرق اكبر الحيوانات حجما على وجه الأرض ليس لديه اسنان ولكنه يندفع بقوة في المياه ليقتات على احياء البحر الطافية فيه.**
37	An Algerian athlete won the gold medal in the world marathon in the midst of cheers from the audience.	**2.41**	**0.90**	**ربح متسابق من الجزائر الميدالية الذهبية في سباق الماراثون العالمي وسط هتاف الجمهور.**
	A player should strive to win the tournament to reward the audience who heartened him.			**من الواجب ان يجتهد اللاعب في الفوز بالبطولة حتى يكافأ الجمهور الذي يشجعه.**
38	Make sure that you live near the university so you will not face any difficulty to go forth.	**1.88**	**0.75**	**احرص على ان تسكن قرب الجامعة حتى لا يصعب عليك الذهاب إليها.**
	I live in a house nearer to the city centre from the place of my work and my children's school.			**اسكن في بيت أقرب الى مركز المدينة من محل عملي ومدارس اولادي.**
39	A person who has money has to pay zakat and give it to the poor, needy, debtors and for God's seek.	**2.32**	**0.66**	**يجب على الرجل ذي المال ان يخرج زكاة ماله ويعطيها للفقراء والمساكين والغارمين وفي سبيل الله.**
	You may be a merchant with a great asset in the bank but you can be stingy to spend it on your family or give to charity from your money.			**قد تكون تاجرا وتملك رصيدا كبيرا في البنك لكنك تكون بخيلا ان لم تنفق على عائلتك او تتصدق من مالك.**
40	Education is the main driver in the development of civilizations and the axis of measuring the evolution and development of communities.	**0.18**	**0.49**	**التعليم هو المحرك الاساسي في تطوير الحضارات ومحور قياس تطور ونماء المجتمعات.**
	The head contains most of the senses enjoyed by humans such as hearing, sight, smell and taste.			**يحتوي الراس على معظم الحواس التي يتمتع بها الانسان كالسمع والبصر والشم والذوق.**
41	Sugar is dissolved in water when adding the right amount with continuous stirring.	**0.03**	**0.18**	**يذاب السكر في الماء عند وضع الكمية المناسبة مع التحريك المستمر.**
	Literacy programs for adults are an important in addition to the march of education in the developing countries.			**برامج محو الأمية للكبار تعد إضافة مهمة إلى مسيرة التعليم في الدول النامية.**

42	The Cobra is the most dangerous snake known to man for its killing venom and it lives in the woods of Africa and India.	**3.47**	**0.61**	**تعتبر الكوبرا اخطر افعى معروفه عند الانسان بسمها القاتل وتعيش في غابات افريقيا والهند.**
	Large numbers of the dangerous venomous Cobra spread over in India, which its poison is considered as the most deadly one and can kill a person within few seconds.			**تنتشر في الهند اعدادا كبيرة من افعى الكوبرا الخطرة التي يعد سمها من أنواع السموم القاتله التي تكفي لقتل الشخص في ثواني.**
43	Did the man with money spend his money on the poor and the needy to gain the approval of God?	**3.81**	**0.42**	**هل انفق الرجل ذو المال على الفقراء والمحتاجين ليكسب رضا الله سبحانه ؟**
	This man is generous and has money and pays the Zakat and spends it on the poor and needy.			**هذا الرجل كريم ويملك مالا وفير يخرج زكاة ماله وينفق منه على الفقراء والمحتاجين.**
44	I went with my children on a trip to France in the summer and Paris was very crowded.	**2.59**	**0.83**	**ذهبت واولادي في رحلة الى فرنسا في الصيف وكانت العاصمة باريس مزدحمة جدا بالسياح.**
	A lot of people prefer to travel to London to attend the Summer Olympics.			**يفضل الكثير من الناس السفر للعاصمة البريطانية لندن لحضور الالعاب الاولمبية الصيفية.**
45	Aldar Al-alamiya publisher in Bahrain published a book entitled A Message to Man across Time.	**1.46**	**0.80**	**صدر كتاب بعنوان رسالة للانسان عبر الزمان من مطبعة الدار العالمية في البحرين.**
	Scientific library in Lebanon has many various and useful books in different fields such as literature, history and scientific facts.			**يوجد في المكتبة العلمية في لبنان الكثير من الكتب المتنوعة والمفيدة في علوم شتى كالادب والتاريخ والحقائق العلمية.**
46	Antidote used to handle poisonous snake bite is to be used only under the supervision of a specialist doctor.	**0**	**0**	**لا يستخدم الترياق لمعالجة اللسعة السامة للافعى الا باشراف طبيب متخصص.**
	Do you want to wear the blue dress in the concert today or you prefer wearing the red one?			**هل تريدين أن تلبسي الفستان الأزرق اليوم في الحفلة أم تفضلين الأحمر ؟**
47	Let’s have a delicacy in that restaurant which is located near our house next to the beach and forget about downtown restaurants.	**1.05**	**0.74**	**لنتاول طعاما شهيا في ذلك المطعم الذي يقع قرب بيتنا المجاور للشاطئ ودعونا من مطاعم مركز المدينة.**
	My mother would like to go shopping from the recently opened stores in our region only because it is closer to our			**والدتي ترغب بالتسوق من المحال التجارية التي افتتحت مؤخرا في منطقتنا فقط لانها اقرب الى بيتنا من مركز المدينة.**

	house than the city centre.			
48	It is pride for every Egyptian to know that Allah has mentioned Egypt in the Quran four times.	**0.06**	**0.25**	**فخر لكل مصري ان يعلم ان الله قد ذكر مصر في القرآن باسمها اربع مرات.**
	Whoever possesses wisdom has owned the lead in managing his own affairs and the affairs of others.			**من ملك الحكمة ملك زمام المبادرة في ادارة شؤونه وشؤون الاخرين.**
49	Immigrants to Canada need to take their winter clothes with them because of the rough weather there.	**3.75**	**0.42**	**ياخذ الاشخاص المهاجرين الى كندا ملابس شتوية سميكة بسبب قسوة المناخ هناك.**
	I read in the book of immigration to Canada that all immigrants have to provide themselves with woollen clothes and shoes lined with fur because it is so cold there.			**قرأت في كتاب الهجرة الى كندا انه على كل مهاجر ان يتزود بملابس صوفيه واحذية مبطنة بالفرو فالبرد قارس فيها.**
50	The Falcon is a member in the group of birds of prey, the longest-lived bird and feeds by hunting rabbits and birds, and is called by many names such as Baz and Bashiq	**3.57**	**0.59**	**الصقر من مجموعة الطيور الجارحة وهو أطول أنواع الطيور عمرا حيث يتغذى على صيد الارانب والطيور ويطلق عليه العديد من الأسماء كالباز والباشق.**
	The Falcon is considered one of the most prominent vultures in the desert and the longest-lived and feeds on hunting the animals.			**يعتبر الصقر من ابرز الكواسر الموجودة في الصحراء واطولها عمرا وهو يقتات على صيد الحيوانات.**
51	Many tourism companies do their best to provide tours to areas associated with ancient civilization such as Petra and the pyramids.	**1.93**	**0.81**	**تعمل العديد من شركات السياحة على توفير رحلات سياحية لمناطق مرتبطة بالحضارات القديمة كالبتراء والأهرامات.**
	Istanbul is the summer capital for many tourism agencies in the Middle East.			**تعد مدينة اسطنبول عاصمة السياحة الصيفية لكثير من مكاتب السياحة في منطقة الشرق الاوسط.**
52	Tohoku earthquake that hit Japan in 2011 is one of the deadliest earthquakes worldwide where the magnitude of 8.9 has cost the country great financial losses.	**1.98**	**0.88**	**يعد زلزال توهوكو الذي ضرب اليابان عام 2011 من أعنف الزلازل عالميا حيث بلغت قوته 8.9 كبد البلاد خسائر مادية كبيرة.**
	I like to travel to Japan, I have heard a lot about its capital but I am afraid that a devastating earthquake hits, just like the one happened in the past year.			**احب السفر الى اليابان سمعت عن عاصمتها الكثير ولكني اخشى ان يضربها زلزالا مدمرا كالذي حدث العام الماضي.**
53	Did you pay a visit to some of the reserves in Africa and watch the lions?	**0.02**	**0.13**	**هل قمت بزيارة لبعض المحميات في افريقيا وشاهدت اسودا فيها ؟**

	Do not drink rotten milk because it could kill you or cause you severe intestinal diseases.			ابتعد عن اللبن الفاسد لأنه قد يقتلك او يتسبب لك بأمراض معوية شديدة.
54	I offer my apologies for the delay in attending the meeting, held in Amman as I could not catch the plane.	0.10	0.30	اقدم اعتذاري عن تأخري في حضور الاجتماع المنعقد في عمان لعدم تمكني من اللحاق بالطائرة.
	I respect my father no matter if he reproaches me for failing to do certain things because he has more experience than I do.			احترم والدي مهما وبخني في امور يشعر اني مقصر فيها فهو اكثر تجربة مني.
55	Make sure to perform you school homework without delay to be delivered on time just to get the best marks.	0.02	0.13	أحرص على اداء واجباتك المدرسية بدون تأخير لتسليمها في الوقت المحدد تماما للحصول على افضل العلامات.
	Whoever provokes hatred among the people has to know that its effects will reach him.			من أثار أسباب الكراهية والحقد بين الناس فعليه أن يعلم أن آثارها السيئة سوف تصله.
56	Satellite channels, being the most important media make an effort to broadcast the news and events moment by moment to make the citizen in the centre of the event and up to date with the latest developments in the world	3.25	0.63	تعمل القنوات الفضائية بوصفها اهم وسائل الاعلام على نقل الاخبار والاحداث لحظة بلحظة تجعل المواطن وسط الحدث مواكبا لتطورات الاحداث في العالم.
	There are many news stations hunting news and events around the world and display them smartly.			توجد العديد من المحطات الأخبارية التي تلاحق الأخبار والاحداث حول العالم وتعرضها بطريقة ذكية.
57	If you have an apple every day that will increase your health and happiness to enjoy the many benefits and thus the doctor will have no a job to do.	3.75	0.43	اذا تناولت تفاحة كل يوم تزيدك صحة فتنعم بالسعادة لفوائدها الكثيرة عندها لن يجد الطبيب عملا له.
	Just one apple a day will make you healthier and may keep you away of the clinic.			تفاحة واحده في اليوم تجعلك اكثر صحة و قد تجنبك زيارة الطبيب طول عمرك .
58	Specialized fitness teams in America and Britain are campaigning to make people aware of the dangers of being overweight.	1.36	0.79	تقوم فرق متخصصة باللياقة البدنية في امريكا وبريطانيا بحملات لتوعية الناس بمخاطر زيادة الوزن.
	Health teams across the country deployed in the eradication malaria completely.			قامت الفرق الصحية المنتشرة في انحاء البلاد بالقضاء على مرض الملاريا تماما.

59	Do you have a book about the administration that include a display of the methods and modern concepts about patterns of leadership?	**1.40**	**0.82**	**هل لديك كتابا عن الادارة يتضمن عرض اساليب ومفاهيم حديثة عن انماط القيادة ؟**
	I read a new book about the education of children in a modern scientific way that encourages the good treatment for them and warns about transferring the unacceptable behaviours to them.			**قرات كتابا جديدا يتحدث عن تربية الاطفال بطريقة علمية يشجع على حسن التعامل معهم ويجنب نقل التصرفات الغير مقبولة لهم.**
60	Due to the intensity of rainfall many workers had to stay hiding under the umbrella.	**0**	**0**	**نظرا لشدة هطول الأمطار فقد بقي العديد من العاملين مختبئين تحت المظلة.**
	The government has sent squad of officers specialized in the field of aviation to France for a training course.			**ارسلت الحكومة كوكبة من الضباط المتخصصين في مجال الطيران الى فرنسا في دورة تدريبية.**
61	A famous wrestler applied to the games organizing committee to participate in the local wrestling championships.	**0.02**	**0.13**	**تقدم فتى المصارعة المشهور بطلب الى لجنة تنظيم المباريات المحلية للمشاركة في بطولة المصارعة.**
	Is your son is afraid of boarding a plane for a long time because he feels nauseous?			**هل يخاف ابنك من ركوب الطائرة لوقت طويل لشعوره بالغثيان ؟**
62	President of the University has honoured the outstanding students in their studies at all Faculties with precious presents.	**0.69**	**0.60**	**أكرم رئيس الجامعة الطلبة المتفوقين في دراستهم في جميع الكليات بهدايا ثمينة.**
	The most generous people for Allah are the ones with most piety and belief and good work.			**ان اكرم الناس عند الله تعالى هم الاكثر تقوى وايمانا وعملا صالحا.**
63	Young people in poor communities suffer harsh childhood because of the deteriorating harsh living conditions.	**3.54**	**0.58**	**يعاني الصغار في المجتمعات الفقيرة من طفولة قاسية بسبب الاوضاع المعيشية المتردية.**
	Children living in poor countries have a difficult life of the weakness of the economy that has forced many of them to work and thus lose their childhood.			**يعيش الاطفال في الدول الفقيرة حياة صعبة بسبب ضعف اقتصادها قد يضطر الكثير منهم الى العمل فيفقد مرحلة الطفولة.**
64	The envier is a person feels inferiority for what the others have and wants it to go away from them and has what they have.	**3.71**	**0.44**	**الحسود هو شخص يشعر بالنقص تجاه مايملكه الاخرين ويرغب في ان تزول عنهم ويملك ما لديهم.**

	A jealous man is the one who does not like goodness for others and wishes the demise of their grace, and we seek refuge with Allah from him.			الإنسان الحسود هو من لا يحب الخير لغيره ويتمنى زوال نعمته وعلينا الاستعاذة بالله منه.
65	My mother does not allow me to leave my room to play till she makes sure I have fully done my school homework	2.07	0.80	لا تسمح والدتي لي بان اغادر غرفتي للعب حتى تتأكد من اني انهيت واجبي المدرسي تماما.
	I work hard in preparation for the review of my classes to get good results in the examinations at the end of the academic year.			اجتهد في مراجعة دروسي تمهيدا للحصول على نتائج جيدة في الامتحانات في نهاية العام الدراسي.
66	Tigris and Euphrates rivers join together in the associated area in the province of Basra to form the Shatt al-Arab.	1.19	0.75	يرتبط نهري دجلة والفرات في منطقة في محافظة البصرة ليشكلا معا شط العرب.
	Strait of Hormuz is linked to the Arabian Gulf on one side, and the Gulf of Oman and the Arabian Sea on the other hand.			يرتبط مضيق هرمز بالخليج العربي من جهة و بخليج عمان وبحر العرب من جهة أخرى.
67	I recommend you to go to the library and to spend a quality time among thoughts of authors.	0.02	0.13	انصحك ان تذهب إلى المكتبة وتقض فيها وقتاً نافعاً بين عقول المؤلفين.
	Gold is one of the important metals in the economic world and has an essential impact on the market movement.			الذهب هو احد المعادن المهمة في عالم الاقتصاد ومحرك اساسي لحركة الاسواق.
68	Do not stop praying in the Masjid especially the dawn prayer for its great reward.	0.10	0.30	لا تترك الصلاة في المسجد وخصوصا الفجر فإن أجرها عظيم وثوابها جزيل.
	Allah sends the apostles as evangelists and warners when evil and injustice grow anywhere on earth.			يرسل الله الرسل مبشرين ومنذرين عندما يكثر الشر ويزداد الظلم في اي مكان في الارض.

Figure 6.4 shows the distribution of the similarity ratings in the full ASTSS-68 dataset. The dataset is well balanced, if one considers that ~ 1/3 of the short text pairs are high, ~ 1/3 low and ~ 1/3 across the broad, difficult medium similarity band from 1.0 - 3.0.

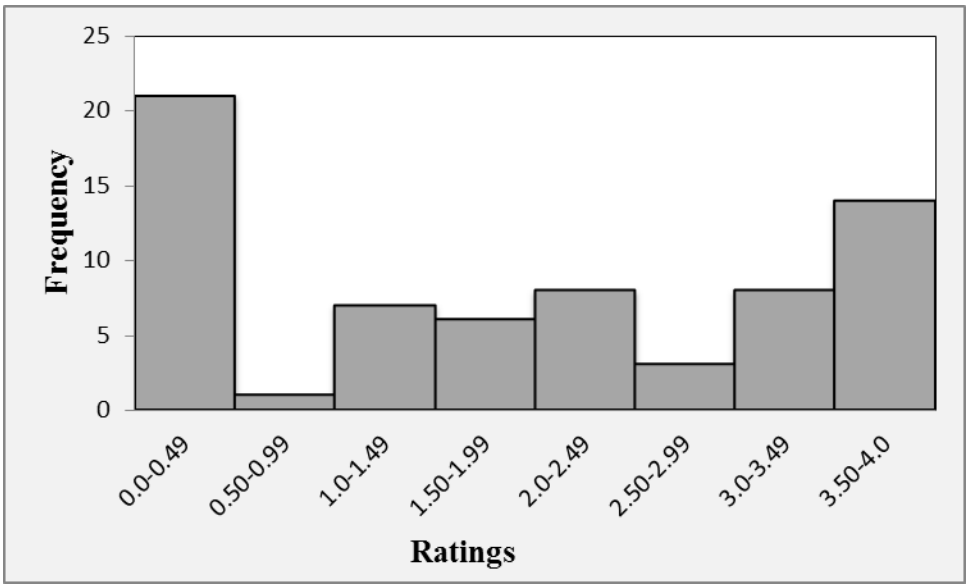


Figure 6.4 Distribution of the similarity ratings in ASTSS-68 dataset.

Prior work in English short text semantic similarity (O'Shea et al., 2013) provided evidence that the card sorting with semantic anchors technique provides ratings that can be legitimately treated as being on a ratio scale (O'Shea et al., 2013). The correlation coefficient (considered in the noun and verb datasets creation procedures) is a suitable statistic that can be applied for measures made on a ratio scale. In this study, the Pearson product moment correlation coefficient was used to identify the consistency of similarity judgments for each participant with the rest of group. This was undertaken using the leave-one-out resampling technique as described in chapter 5 (section 5.2.4) whereby the correlation coefficient for each of the 62 participants was calculated between the participant's ratings and the average ratings of the rest of group. Figure 6.5 shows the correlation coefficients of 62 Arabic participants on the ASTSS-68 dataset.

The possible indicative value and bounds of performance expected from a computational Arabic short text similarity algorithm attempting to perform the same task have been calculated as the average, worst and best performances of human

participants on the ASTSS-68 dataset as shown in table 6.13. Whereby, if any Arabic machine algorithm equals or exceeds the average of the correlations of all participants (r = 0.892), it will be considered performing well. The worst performing participant of (r = 0.80) is considered as the lower bound for the expected performance whereas any machine algorithm coming close to the best performing participant at 0.970 would be considered as performing very well.

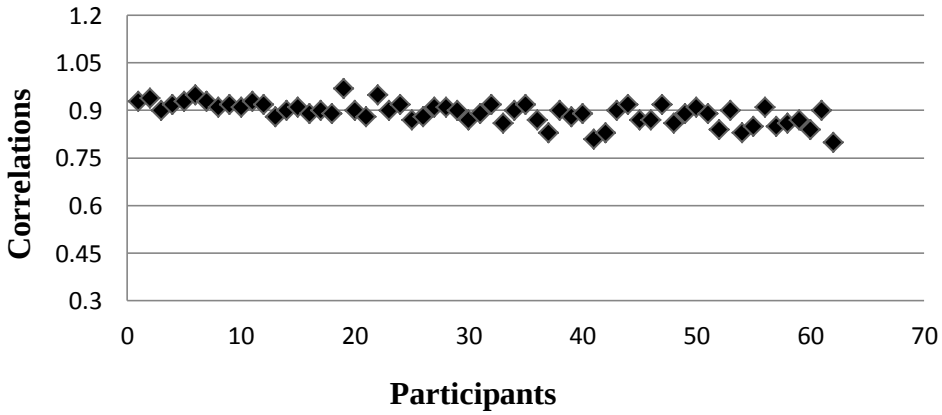


Figure 6.5 The Correlation Coefficients of 62 Arabic Participants

Table 6.13 The Correlation Coefficient with Mean Human Judgements

	Correlation (*r*)
Average of the correlation of all participants	0.892
Best participants	0.970
Worst participants	0.80

6.3 Evaluation of the Arabic Short Text Sematic Similarity (NasTa) Framework

The development process of the NasTa framework consisted of two phases as described in chapter 4. The first phase concerned the creation of the NasTa-A which focused on the noun semantic similarity whilst the second, NasTa-F was created based on the Part of Speech (POS) and word sense disambiguation. The Arabic short text benchmark (ASTSS-68) dataset created in this chapter was used to assess the

accuracy of NasTa-A and NasTa-F. This allowed the determination of which combination should be used profitably in NasTa framework by means of comparing the performance of the NasTa-A and NasTa-F. The evaluation process of NasTa consisted of three major stages which included:

1. Creation of an optimization dataset in order to determine the optimal parameter values of the NasTa.
2. Evaluation of the NasTa-A using ASTSS-68 dataset.
3. Evaluation of the NasTa-F using ASTSS-68 dataset.

6.3.1 Creation of an Optimization Short Text Pairs Set

A set of 21 Arabic Short Text Semantic Similarity (ASTSS-21) pairs was produced in order to use it to optimize the NasTa parameters process. This set was created using the rest of 65 short text pairs used in the ASTSS-68 dataset to select the final set of the short text pairs, section (6.2.2.2). The set of 21 short text pairs consisted of 7 high similarity short text pairs which were selected from the rest of 29 the short text pairs rated high by participants in the experiment to select the high similarity pairs for the ASTSS-68 dataset and 7 medium similarity pairs were selected from the rest of 36 pairs rated medium by participants in the experiment. Seven low similarity short text pairs were selected randomly from the database of 1088 Arabic short texts created in section (6.2.2.1). Looking at the 7 lowest similarity pairs in ASTSS-68 dataset, they are all either 0 or very close to 0 (appendix 8). In every case where a rating is non-zero, the SD is substantially higher than the rating itself, implying that all of these ratings are effectively 0 with an element of noise superimposed. Therefore, the decision was made to simply allocate the value 0 to the 7 low similarity pairs for ASTSS-21. The set of 21 short text pairs with human ratings is presented in appendix 9.

6.3.2 Evaluation of the NasTa-A

This section describes the evaluation process of the NasTa-A algorithm which calculates the similarity by combining the noun semantic similarity and word order similarity of the compared short texts. The evaluation process has two aims. Firstly,

to identify the quality of NasTa-A by means of an investigation of its performance compared with human perception using the ASTSS-68 dataset. Secondly, to investigate the influence of word order similarity in the NasTa-A performance, whereby the MSA is considered syntactically flexible, i.e. it has a relatively free word order, as described in chapter 4.

6.3.2.1 Evaluation's Methodology

The evaluation methodology consisted of two steps. These are the determination of the optimal parameter values of NasTa-A algorithm and the application of the ASTSS-68 dataset pairs on the NasTa-A algorithm with and without the word order similarity component.

For the first step, NasTa-A requires determining the optimal values for three parameters before use. These are a threshold for the semantic vector derivation, a threshold for the word order vector formation and **δ** for adjusting the relative contributions of semantic and word order information to the final NasTa-A calculation. At this stage, it was decided to use the values used by Li et al. (2006) in the initial testing experiment. A value of 0.2 was used for the semantic threshold, 0.4 was used for the word order threshold and a value of 0.85 was used for **δ**. These parameter values were determined using a small set of short text pairs (Li et al., 2006). Furthermore, the Arabic Noun Semantic Similarity (KalTa-A) measure with its pre-determined optimal parameters values ($\alpha = 0.12$ and $\beta = 0.21$) was used to calculate the similarity between the nouns in both short texts as decided in chapter 5.

In the second step of the evaluation process, the short text pairs on the ASTSS-68 were run using the NasTa-A with and without the word order similarity component in order to produce the machine similarity ratings in the range from 0 to 1.

6.3.2.2 Evaluation's Results

The results of the evaluation process are presented in table 6.14 which shows the human similarity ratings with the corresponding machine similarity ratings produced by NasTa-A with and without the word order on the ASTSS-68 dataset. In table 6.14, the second column represents the human similarity ratings which were rescaled from

0 - 4 to 0 – 1 for the purpose of comparison. The third column represents the machine similarity ratings produced by the NasTa-A with the word order component whilst the final column represents the machine similarity ratings generated by the NasTa-A without (WO) the word order similarity.

Table 6.14 Short Text Similarity Ratings for ASTSS-68 dataset from Human and NasTa-A.

ST Pairs	Human Ratings	NasTa-A Ratings	NasTa-A without WO Ratings	ST Pairs	Human Ratings	NasTa-A Ratings	NasTa-A without WO Ratings
1	0.95	0.54	0.56	35	0.97	0.90	0.97
2	0.85	0.73	0.78	36	0.93	0.58	0.62
3	0.53	0.45	0.45	37	0.60	0.26	0.27
4	0.42	0.34	0.33	38	0.47	0.53	0.58
5	0.00	0.22	0.24	39	0.58	0.20	0.19
6	0.02	0.19	0.17	40	0.05	0.19	0.22
7	0.86	0.68	0.74	41	0.01	0.32	0.32
8	0.90	0.75	0.79	42	0.87	0.75	0.82
9	0.95	0.61	0.68	43	0.95	0.71	0.70
10	0.60	0.35	0.38	44	0.65	0.57	0.59
11	0.33	0.30	0.27	45	0.37	0.35	0.38
12	0.01	0.14	0.13	46	0.00	0.04	0.05
13	0.06	0.22	0.20	47	0.26	0.34	0.32
14	0.96	0.53	0.57	48	0.02	0.27	0.28
15	0.83	0.69	0.71	49	0.94	0.54	0.57
16	0.89	0.25	0.25	50	0.89	0.61	0.65
17	0.59	0.40	0.42	51	0.48	0.51	0.55
18	0.36	0.56	0.58	52	0.50	0.40	0.45
19	0.02	0.12	0.13	53	0.01	0.00	0.00
20	0.01	0.08	0.07	54	0.03	0.14	0.12
21	0.75	0.38	0.41	55	0.01	0.38	0.38
22	0.81	0.66	0.71	56	0.81	0.72	0.75
23	0.53	0.45	0.48	57	0.94	0.37	0.38
24	0.63	0.44	0.43	58	0.34	0.28	0.30
25	0.43	0.45	0.47	59	0.35	0.44	0.47
26	0.00	0.16	0.16	60	0.00	0.21	0.22
27	0.01	0.28	0.29	61	0.01	0.17	0.19
28	0.84	0.53	0.53	62	0.17	0.17	0.20
29	0.88	0.62	0.68	63	0.89	0.40	0.40
30	0.84	0.46	0.47	64	0.93	0.45	0.50
31	0.56	0.43	0.46	65	0.52	0.21	0.24
32	0.45	0.27	0.28	66	0.30	0.50	0.51
33	0.05	0.11	0.10	67	0.01	0.44	0.46
34	0.01	0.17	0.19	68	0.03	0.04	0.04

6.3.2.3 Discussion

The value of NasTa-A is assessed by computing the correlation coefficient between the average ratings of human participants on the ASTSS-68 dataset and the machine ratings obtained from NasTa-A. The Pearson product-moment correlations (r) for NasTa-A (with WO) and NasTa-A without WO are presented in table 6.15. The results in table 6.15 illustrate that the NasTa-A at (r = 0.785) performs significantly below the average of the correlation of human performance at (r = 0.892). Result from one sample t-test which was used to compare between a single correlation (NasTa-A) and the average of the correlation coefficients on the ASTSS-68.

Null hypothesis (H_0) is to test of $\mu = 0.785$ vs. $\mu \neq 0.785$. The result of the one sample t-test with confidence interval plot is summarized in the figure 6.6. The true mean could lie anywhere in the interval (0.883, 0.901), the sample mean (n=62) is 0.892 and t-test statistic is 24.45 with P-value < 0.0001. Since the P-value is less than the significance level (0.05), the null hypothesis can be rejected.

Table 6.15 The Performance of NasTa-A on the ASTSS-68 dataset.

On ASTSS-68 Data Set	Correlation *r*
NasTa-A algorithm	0.785
NasTa-A without WO algorithm	0.786
Average of the correlation of all participants	0.892
Best participants	0.970
Worst participants	0.80

Also the NasTa-A without WO at (r = 0.786) performs significantly below the average of the correlation of human performance at (r = 0.892) with P-value < 0.0001. Furthermore, the results in table 6.15 illustrate that the performance of the NasTa-A at (r = 0.785) was below the worst human (lower bound) performance at (r = 0.80).

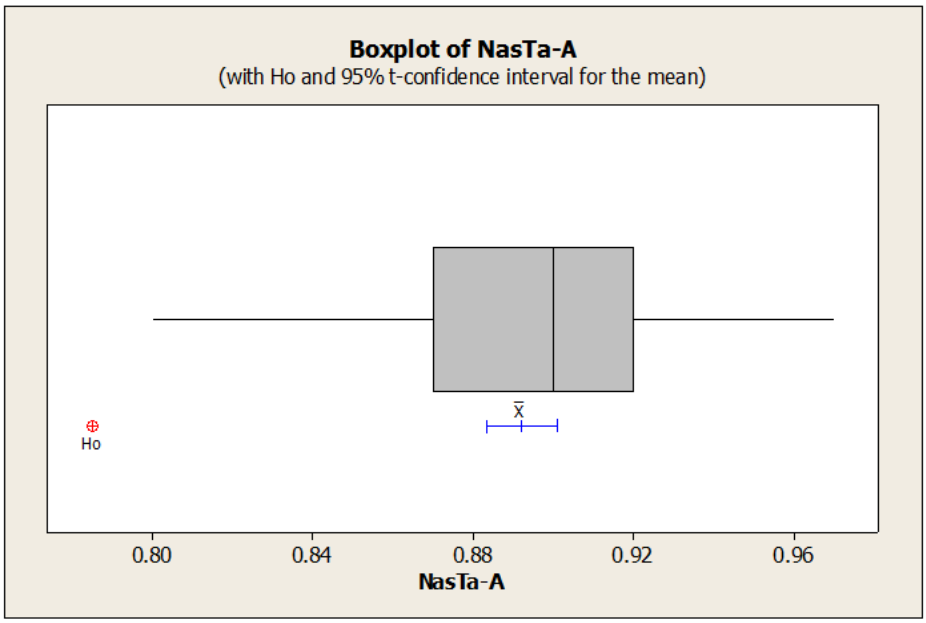


Figure 6.6 Results for the one sample t-test

Steiger's z-test was used to compare the difference between NasTa-A and NasTa-A without WO in order to investigate the influence of the word order similarity in the NasTa-A performance. Using Steiger's z-test requires the construction of a correlation triangle (3 correlations) between:

NasTa-A ratings vs. Human ratings = 0.785
NasTa-A without WO vs. Human ratings = 0.786
NasTa-A vs. NasTa-A without WO = 0.996

n = 68 (the number of short text pairs in the ASTSS-68 dataset)

Applying the test (using the online calculator which was available at (Grabin, 2013)) indicates that the difference between NasTa-A and NasTa-A without WO is not statistically significant (Z = -0.15, p = 0.878). This result also indicates that the word order similarity has no influence on the performance of the NasTa-A.

Figure 6.7 shows the correlation between the NasTa-A and human ratings on the ASTSS-68 dataset. The NasTa-A has not performed as well as might be expected, failing to give similarity values close to human ratings for many short text pairs in each similarity range (low to high) as shown in figure 6.7. For example, the short text

pairs (22, 42, 70, 81, 90 and 91) rated high by participants but obtained low similarity or low medium similarity values by NasTa-A as shown in table 6.14.

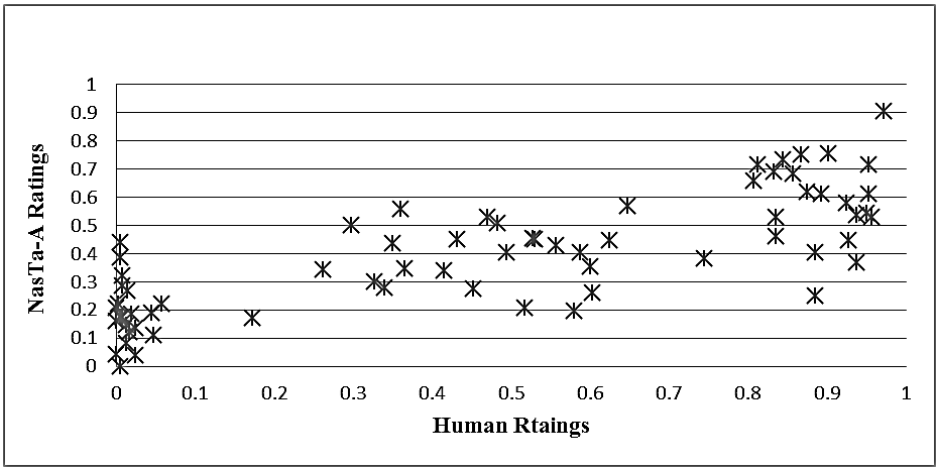


Figure 6.7 The Correlation between the Human ratings and NasTa-A measure

Furthermore, Steiger's z-test showed that the word order component has no influence on the performance of the NasTa-A. It will be born in mind that the NasTa-A parameters were set using values determined for English and this might have led to the unexpected performance. Therefore an experiment was performed to investigate optimising the NasTa-A parameters to see if NasTa-A could be improved. This experiment is described in section 6.3.2.4.

6.3.2.4 Optimising Parameters Experiment

The set of 21 Arabic short text pairs (ASTSS-21) created in section (6.3.1) was used in the parameter optimization experiment. As described in chapter 4, the overall short text semantic similarity of the NasTa-A calculated using the following formula.

$$S(T_1, T_2) = \delta * \text{Semantic similarity} + (1 - \delta) * \text{word order similarity} \qquad (6.1)$$

Where $0.50 < \delta \leq 1$, the syntax (word order similarity) plays a subordinate role for the semantic text processing (Wiemer-Hastings, 2000) therefore Li et al. (2006) proposed that the value of δ parameter should be greater than 0.50.

Two aspects were necessary for consideration as regards the semantic threshold: the detection and utilisation of the similar semantic features of words to the greatest extent and the maintenance of low noise. It was necessary to use an appropriately small semantic threshold in order to permit the model to obtain adequate semantic information distributed across every word. Where the threshold had too low a value, excessive noise to the short text similarity measure arose, resulting in deterioration of the overall performance of the measure. Consequently, the initial value given to the semantic threshold parameter was 0.20. This consideration also applied to the word order similarity threshold, thus a higher value was utilised for this. It is necessary for a pair of linked words (the most similar in two short texts) to be intuitively relatively similar in order that the word order vector can be used as, if this does not apply, the relative ordering of pairs of words with less similarity offers very little information. The initial value given to the word order threshold parameter is 0.30.

Given the initial value of each parameter, the short text pairs on the ASTSS-21set were run using the NasTa-A to produce machine similarity ratings in a range of 0 to 1. The correlation coefficient between the human ratings of ASTSS-21set and those obtained from the NasTa-A was computed. The values of the Arabic measure parameters were changed to obtain a set of correlation coefficients. The increasing step of each parameter was 0.05. Then the parameters with the strongest correlation coefficient were considered as the optimal parameters. In this experiment, the strongest correlation coefficient was obtained at $\delta = 1$ and the semantic threshold = 0.2.

Using the identified value of δ parameter with the formula 6.1, the value of the word order similarity component will equal to 0. This result indicates that the word order component has no influence in the NasTa-A performance which confirms the results obtained from the evaluation process of the NasTa-A. The ratings produced by NasTa-A using the new parameter values is the same rating produced by NasTa-A without WO in table 6.14. The correlation coefficient between the NasTa-A ratings and the human ratings is 0.786 which is below the average human performance of 0.892.

The performance of the NasTa-A was affected negatively by two issues. NasTa-A focused only on the similarity of nouns and ignores other Parts of Speech (POS). In addition, the NasTa-A relied largely on computing the similarity between the nouns in both short texts but did not take the context in which they occur into account and thus affects the final short text similarity score. To illustrate these issues, the following short text pair selected from the ASTSS-68 dataset (pair number 67) offers an example.

T_1: انصحك ان تذهب إلى المكتبة وتقض فيها وقتاً نافعاً بين عقول المؤلفين

I recommend you to go to the library and to spend a quality time among thoughts of authors.

T_2: الذهب هو احد المعادن المهمة في عالم الاقتصاد ومحرك اساسي لحركة الاسواق

Gold is one of the important metals in the economic world and has an essential impact on the market movement.

Step 1 is to transform the two short texts to Verb-Subject-Object order. The short text T_1 already has a VSO order and T_2 is an equational (verbless) short text. As described in chapter 2, the equational sentence is a sentence without a verb and its structure consists of the subject and predicate.

Step 2 is to create the joint word set T for the short text T_1 and the short text T_2:

{انصحك, ان, تذهب, الى, المكتبة,و تقضي, فيها, وقتا, نافعا, بين, عقول, المؤلفين, الذهب, هو, احد, المعادن, المهمة, في, عالم, الاقتصاد, ومحرك, اساسي, لحركة, الاسواق}

Step 3 involves the calculation of the semantic similarity component (S_S), where the semantic vectors for the two short texts T_1 and T_2 can be created from the joint word set T and corpus statistics. Table 6.16 illustrates the process of the creation of the semantic vector for T_1. The first rightmost column in table 6.16 lists words in T whilst the first row lists words in the short text T_1. The words in the first column and row are listed in the order as they occur in the joint word set T and the short text T_1.

Table 6.16 The Semantic Vector Creation Process.

		T_1 (w_i)												
Weight $I(w_i)$. $I(\hat{w}_i)$	S	12.المؤلفين authors	11. عقول thoughts of	10.بين among	9. نافعا a quality	8. وقتًا time	7. فيها in it	6. ونقضي to spend	5.المكتبة the library	4. الى to	3. تذهب go	2. ان to	1. انصحك I recommend you	Joint word set $(\hat{w}_i)$
0.260	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	1.انصحك I recommend you
0.025	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	2. ان to
0.489	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	3. تذهب go
0.078	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	4. الى to
0.227	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	5. المكتبة the library
0.394	1.000	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	6. ونقضي to spend
0.025	1.000	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	7. فيها in it
0.376	1.000	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8. وقتًا time
0.321	1.000	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9. نافعا a quality
0.074	1.000	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10. بين among
0.197	1.000	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11. عقول thoughts of
0.212	1.000	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12. المؤلفين authors
0.293	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	**13. الذهب gold**
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	14. هو is
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15. احد one of
0.422	0.634	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.634	0.00	0.00	**16. المعادن metals**
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17. المهمه the important
0.025	1.000	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	18. في in
0.151	0.536	0.536	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**19. عالم world**
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20. الاقتصاد economic
0.404	0.215	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.215	0.00	0.00	0.00	0.00	21. ومحرك impact
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22. اساسي an essential
0.207	0.393	0.00	0.00	0.00	0.00	0.393	0.00	0.00	0.00	0.00	0.00	0.00	0.348	23. لحركة movement
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24. الاسواق the markets

With regard to each word in the joint word set T, the cross point cell must be set to 1 if the same word occurs in the short text T1. If this is not undertaken, the cross point cell of the most similar word should be set at its similarity value or 0, reliant on whether the highest similarity score exceeds the threshold of 0.2. For example, the word لحركة "movement" is not in T1, but the most similar word is وقت "time", with a similarity of 0.393. Thus, the cell at the cross point of movement and time is set to 0.393 as it exceeds the threshold of 0.2 and all other words are set to 0. The largest value in each row is chosen to create the lexical vector s_1 for the T_1. The leftmost column lists the corresponding information content I(*w*) to weight the significance of the word. Where each entry value of the lexical vector s_1 is weighted according to the information content of w_i (a word in the joint word set T) and $\hat{w}_i$ (the associated word in the short text T_1 that have the highest similarity score with w_i). For this example, the information content of w_{23} (لحركة "movement") in T is $I(w_{23}) = 0.614$

whilst for $\hat{w}_8$ (وقت "time") in T_1 is $I(\hat{w}_8) = 0.338$, where $I(w_{23}) * I(\hat{w}_8) = 0.207$. The lexical vector cell $s_1(23) = 0.393$ which is weighted by 0.393 * 0.207 = **0.081**. Consequently, the semantic vector for the short text T_1 is:

$\mathbf{s_1}$: {0.260, 0.025, 0.489, 0.078, 0.227, 0.394, 0.025, 0.376, 0.321, 0.074, 0.197, 0.212, 0.293, 0.0, 0.0, 0.267, 0.0, 0.025, 0.081, 0.0, 0.087, 0.0, **0.081**, 0.0}.

In accordance with the same process, the semantic vector for the short text T_2 is:
$\mathbf{s_2}$: {0.059, 0.0, 0.293, 0.0, 0.036, 0.0, 0.025, 0.081, 0.0, 0.0, 0.046, 0.081, 0.179, 0.081, 0.144, 0.363, 0.154, 0.025, 0.107, 0.149, 0.716, 0.177, 0.113, 0.138}.

Using s_1 and s_2, the semantic similarity between T_1 and T_2 is $S_S = 0.463$.

Step 4 includes the calculation of the word order similarity component (S_r). The word order vectors were similarly derived, the word order threshold was set to 0.4.

$\mathbf{r_1}$: {1 2 3 4 5 6 7 8 9 10 11 12 3 0 0 3 0 7 12 0 0 0 0 0}
$\mathbf{r_2}$: {0 0 1 0 0 0 6 0 0 0 0 7 1 2 3 4 5 6 7 8 9 10 11 12}

Using r_1 and r_2, the word order similarity is $S_r = 0.297$
Finally, the overall semantic similarity between T_1 and T_2 is 0.44.

This pair of short texts was rated very low (unrelated in meaning 0.02) by participants as shown in table 6.14 whilst the NasTa-A gave it a medium similarity value (0.44). An explanation is provided through looking at the table 6.16. As can be observed, the cell at the cross point of the verb "go" in the short text T_1 and the noun "gold" in the joint word set T is set to 1 (high similarity value). The reason of that the verb "go" and the noun "gold" have the same form which is **ذهب**. The NasTa-A ignores the POS and considers (**ذهب** "go" and **ذهب** "gold") as the same word which gives a high similarity value between the compared words. For the same reason (consider the verb ذهب "go" as a noun by the NasTa-A which mean a "gold"), the cell at the cross point of the verb "go" in the short text T_1 and the noun "metals" in the joint word set T is set to 0.634 (high medium similarity value).

Furthermore, the cell at the cross point of the word مؤلفين "authors" in the short text T_1 and the word عالم "world" in the joint word set T is set to 0.536 (high medium

similarity value). Due to missing diacritics described in chapter4, the Arabic word عالم offers multiple meanings which mean عالَم ***Aalam*** "world" or عالِم ***Aalim*** "scientist". The NasTa-A relied largely on computing the similarity between the nouns in both short texts but did not take the context in which they occur into account. In this case, the comparison between the word مؤلفين "authors" and the word عالم as a "scientist" gave a high medium similarity value and thus affects the final short text similarity score which gave a similarity value far from human ratings.

These two issues affected the performance of the NasTa-A which obtained a correlation significantly below the average of the correlation of human performance on the ASTSS-68 dataset.

6.3.3 Evaluation of the NasTa-F

This section describes the evaluation process of the NasTa-F algorithm which was created to address the weakness of the NasTa-A algorithm (understanding context within a short text structure and the use of POS rather than nouns) described in section 6.3.2. Experimental results in section 6.3.2 offered evidence that the presence of word order similarity has no influence on the performance of the NasTa-A algorithm. The initial decision was to remove the word order similarity component from the NasTa-F algorithm; however calculation of the short text similarity based on POS and WSD may enhance the performance of word order component and thus enhance the overall performance of NasTa-F algorithm. In this case, the evaluation process has three aims including:

1. Identification of the quality of NasTa-F by means of an investigation of its performance compared with human perception using the ASTSS-68 dataset.

2. Investigation of the influence of word order similarity in the NasTa-F via comparing its performance with and without the word order component.

3. Determination of whether a combination should be used profitably in NasTa framework by means of comparing the performance of NasTa-A and NasTa-F.

6.3.3.1 Evaluation Methodology and Results

The evaluation methodology involved the determination of the optimal parameter values of the NasTa-F algorithm and the application of the ASTSS-68 dataset pairs on the NasTa-F algorithm without the word order similarity component. In this case, the short text similarity was calculated based on POS, WSD and semantic similarity.

Chapter 4 presented a new WSD algorithm, namely AWSAD, which was created to disambiguate all the words in the Arabic short text in order to improve the NasTa performance. As described in chapter 2, the evaluation process of the WSD algorithm performance requires Arabic manually sense-tagged corpora. There is no machine method to automate sense-tagging in an Arabic corpus and human sense-tagging is labour intensive. It requires a human expert to be very familiar with each Arabic word's definition. Diab et al. (2007) presented an Arabic all-words sense annotated set in running text but it was not available from the authors for the purpose of research. Consequently, it was decided to evaluate the AWSAD algorithm indirectly with respect to its performance within the NasTa-F algorithm. This was feasible because every other component of NasTa-F had been evaluated in isolation as well as within NasTa-F, avoiding confounding factors. The main idea behind this evaluation method is: the success rate of NasTa-F should increase as the AWSAD algorithm performance gets better.

The first step of the evaluation methodology of the NasTa-F algorithm is to select its parameters' values. For the AWSAD algorithm, the important parameter is the window size. Two different kinds of benefit can be acquired by adjusting the size of the context window. Selection of a large window, (for example, more than five words where the target word in the middle), means more words will be considered to each sense of the target word, thus increasing the likelihood of ascertaining a sense of the target word which bears a close relationship to one or more context window word senses. However, where a small size window is in existence (e.g., three words only, one on each side of the target word), the outcome arises that very few words can be considered for each target word sense. As a result, it is to be expected that the algorithm will locate more appropriate matches. Words closer to the target word are more likely to be related than those which are further from the target word, thus

usage of a small context window may well result in fewer irrelevant words being used. Also the use of a small window results in the WSD algorithm running much faster as fewer comparisons are made. The window sizes used in this section are 3, 5 and 7.

The second parameter of the NasTa-F is the semantic threshold for the semantic vector derivation. The value selected for this parameter is 0.2 was identified in the NasTa-A evaluation process (section 6.3.2). Also, the Arabic noun (KalTa-A) measure was used to calculate the similarity between pairs of nouns and the Arabic verb (KalTa-F) measure with its pre-determined optimal parameter values ($\alpha = 0.2$ and $\beta = 0.459$) was used to calculate the similarity between two verbs. These optimal values were established in chapter 5.

As described in chapter 2, the Arabic content words were classified by traditional Arabic linguistics into verbs and nouns (including adjectives and adverbs) whilst modern linguistics classifies the content words into nouns, verbs, adjectives and adverbs. In this section the AWSAD algorithm is performed based on the two different classification methods in order to investigate the influence of each classification on the performance.

In the second step of the evaluation methodology, the short text pairs on the ASTSS-68 dataset were run using the NasTa-F algorithm in order to produce the machine similarity ratings in the range from 0 to 1. The machine similarity ratings were produced for the window sizes 3, 5 and 7 and based on the modern classification of POS. Table 6.17 shows the human similarity ratings with the corresponding machine similarity ratings produced by NasTa-F on the ASTSS-68 dataset. In table 6.17, the second column represents the human similarity ratings which were rescaled from 0 - 4 to 0 – 1 for the purpose of comparison. The third column represents the machine ratings produced using the window size 3 whilst the last two columns represent the machine ratings generated using the window sizes 5 and 7 respectively (MR means the machine similarity ratings).

In accordance with the same procedure, the machine similarity ratings were produced for the window sizes 3, 5 and 7 and based on the traditional classification of POS. The results of this experiment are presented in table 6.18.

Table 6.17 Short Text Similarity Ratings for ASTSS-68 dataset from Human and NasTa-F without word order based on **Modern classification** with different window sizes.

ST Pairs	**Human Ratings**	**MR Size 3**	**MR Size 5**	**MR Size 7**	ST Pairs	**Human Ratings**	**MR Size 3**	**MR Size 5**	**MR Size 7**
1	0.95	0.66	0.66	0.66	35	0.97	0.90	0.88	0.88
2	0.85	0.76	0.73	0.73	36	0.93	0.58	0.51	0.33
3	0.53	0.31	0.28	0.28	37	0.60	0.24	0.24	0.24
4	0.42	0.26	0.25	0.24	38	0.47	0.41	0.37	0.33
5	0.00	0.19	0.19	0.19	39	0.58	0.53	0.53	0.50
6	0.02	0.06	0.02	0.02	40	0.05	0.20	0.20	0.19
7	0.86	0.69	0.67	0.68	41	0.01	0.22	0.07	0.07
8	0.90	0.76	0.75	0.75	42	0.87	0.75	0.68	0.66
9	0.95	0.68	0.68	0.68	43	0.95	0.75	0.75	0.62
10	0.60	0.41	0.42	0.42	44	0.65	0.52	0.24	0.22
11	0.33	0.34	0.34	0.34	45	0.37	0.31	0.28	0.28
12	0.01	0.11	0.11	0.11	46	0.00	0.12	0.06	0.06
13	0.06	0.20	0.18	0.20	47	0.26	0.37	0.37	0.38
14	0.96	0.80	0.56	0.56	48	0.02	0.11	0.13	0.11
15	0.83	0.68	0.60	0.68	49	0.94	0.58	0.58	0.58
16	0.89	0.72	0.72	0.72	50	0.89	0.80	0.80	0.78
17	0.59	0.52	0.53	0.48	51	0.48	0.52	0.48	0.51
18	0.36	0.63	0.48	0.46	52	0.50	0.4	0.4	0.28
19	0.02	0.10	0.10	0.10	53	0.01	0.08	0.00	0.00
20	0.01	0.03	0.03	0.04	54	0.03	0.31	0.28	0.31
21	0.75	0.58	0.47	0.53	55	0.01	0.01	0.01	0.01
22	0.81	0.70	0.71	0.69	56	0.81	0.58	0.65	0.66
23	0.53	0.60	0.62	0.59	57	0.94	0.4	0.38	0.38
24	0.63	0.63	0.58	0.59	58	0.34	0.14	0.14	0.14
25	0.43	0.41	0.32	0.32	59	0.35	0.33	0.34	0.33
26	0.00	0.13	0.02	0.02	60	0.00	0.13	0.11	0.11
27	0.01	0.14	0.14	0.05	61	0.01	0.15	0.13	0.13
28	0.84	0.67	0.67	0.67	62	0.17	0.17	0.17	0.17
29	0.88	0.72	0.29	0.29	63	0.89	0.59	0.58	0.57
30	0.84	0.56	0.56	0.62	64	0.93	0.66	0.67	0.66
31	0.56	0.55	0.55	0.55	65	0.52	0.29	0.36	0.36
32	0.45	0.11	0.14	0.12	66	0.30	0.54	0.50	0.50
33	0.05	0.10	0.10	0.10	67	0.01	0.10	0.11	0.10
34	0.01	0.13	0.13	0.13	68	0.03	0.12	0.12	0.12

Table 6.18 Short Text Similarity Ratings for ASTSS-68 dataset from Human and NasTa-F without word order based on **Traditional classification** with different window sizes.

ST Pairs	Human Ratings	MR Size 3	MR Size 5	MR Size 7	ST Pairs	Human Ratings	MR Size 3	MR Size 5	MR Size 7
1	0.95	0.66	0.66	0.66	35	0.97	0.9	0.88	0.88
2	0.85	0.76	0.73	0.73	36	0.93	0.58	0.51	0.33
3	0.53	0.31	0.28	0.28	37	0.60	0.24	0.24	0.24
4	0.42	0.26	0.25	0.24	38	0.47	0.56	0.37	0.37
5	0.00	0.21	0.19	0.19	39	0.58	0.53	0.53	0.5
6	0.02	0.05	0.02	0.02	40	0.05	0.2	0.2	0.19
7	0.86	0.69	0.66	0.65	41	0.01	0.22	0.07	0.07
8	0.90	0.76	0.75	0.75	42	0.87	0.67	0.68	0.68
9	0.95	0.68	0.68	0.68	43	0.95	0.75	0.75	0.62
10	0.60	0.42	0.43	0.43	44	0.65	0.52	0.24	0.22
11	0.33	0.34	0.34	0.34	45	0.37	0.31	0.28	0.28
12	0.01	0.11	0.11	0.11	46	0.00	0.12	0.06	0.06
13	0.06	0.2	0.18	0.2	47	0.26	0.37	0.37	0.38
14	0.96	0.77	0.56	0.72	48	0.02	0.11	0.11	0.11
15	0.83	0.68	0.6	0.68	49	0.94	0.58	0.58	0.58
16	0.89	0.72	0.72	0.72	50	0.89	0.8	0.8	0.78
17	0.59	0.58	0.4	0.48	51	0.48	0.52	0.48	0.51
18	0.36	0.66	0.48	0.46	52	0.50	0.4	0.4	0.28
19	0.02	0.1	0.1	0.1	53	0.01	0.08	0	0
20	0.01	0.03	0.03	0.04	54	0.03	0.28	0.32	0.31
21	0.75	0.51	0.53	0.53	55	0.01	0.01	0.01	0.01
22	0.81	0.7	0.71	0.69	56	0.81	0.62	0.66	0.65
23	0.53	0.56	0.62	0.54	57	0.94	0.39	0.38	0.4
24	0.63	0.63	0.58	0.59	58	0.34	0.14	0.14	0.14
25	0.43	0.34	0.34	0.34	59	0.35	0.34	0.29	0.32
26	0.00	0.13	0.02	0.02	60	0.00	0.13	0.13	0.11
27	0.01	0.14	0.14	0.05	61	0.01	0.15	0.15	0.13
28	0.84	0.67	0.67	0.67	62	0.17	0.17	0.17	0.17
29	0.88	0.72	0.29	0.29	63	0.89	0.59	0.58	0.57
30	0.84	0.56	0.62	0.56	64	0.93	0.66	0.66	0.67
31	0.56	0.55	0.55	0.55	65	0.52	0.29	0.36	0.36
32	0.45	0.11	0.14	0.12	66	0.30	0.54	0.5	0.5
33	0.05	0.1	0.1	0.1	67	0.01	0.09	0.11	0.09
34	0.01	0.13	0.13	0.13	68	0.03	0.12	0.12	0.12

6.3.3.2 Discussion

The consistency of NasTa-F algorithm with human perception was identified by computing the correlation coefficient between the average ratings of human participants on the ASTSS-68 dataset and the machine ratings obtained from the

NasTa-F for each window size as shown in Table 6.19. Figure 6.8 shows the performance of NasTa-F with respect to different POS classifications and different window sizes. Also figure 6.8 compares between the performances of NasTa-F algorithm with the average of human participants.

Table 6.19 The Performance of NasTa-F without word order on the ASTSS-68 dataset.

On the ASTSS-68 dataset	**Correlation *r***	**Comments**
Average of the correlation of all participants	0.892	
Best participants	0.970	
NasTa-F algorithm / modern POS classification	0.901	Window size 3
	0.883	Window size 5
	0.869	Window size 7
NasTa-F algorithm / traditional POS classification	0.897	Window size 3
	0.882	Window size 5
	0.875	Window size 7

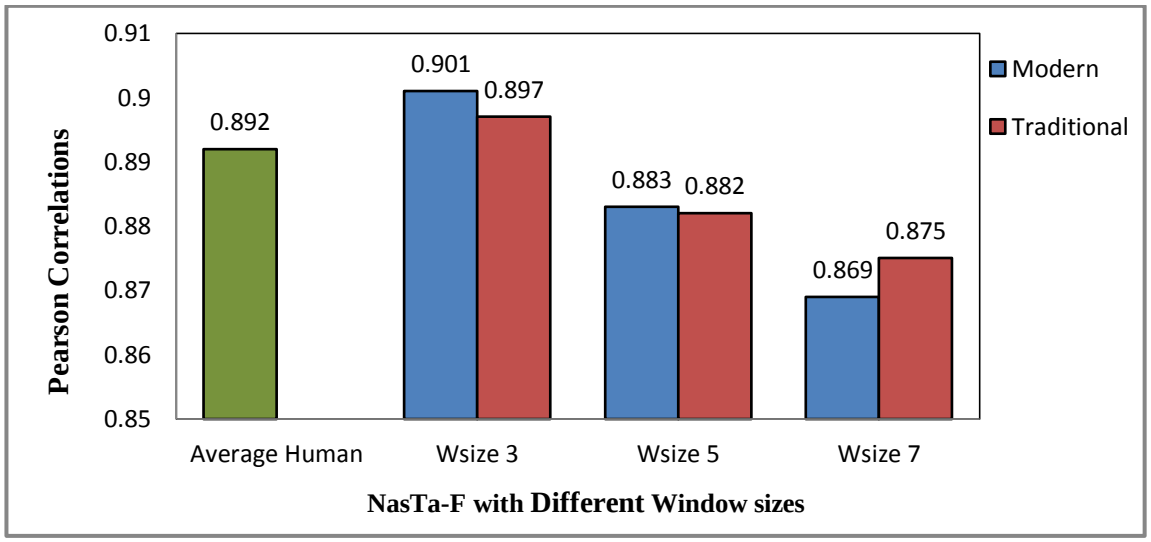


Figure 6.8 the performance of NasTa-F without word order vs. different POS classification and different window sizes.

Figure 6.8 indicates that the performance of the NasTa-F algorithm with window size 3 achieved good correlations with the human ratings for both modern and traditional POS classifications which obtained correlations 0.901 and 0.897 respectively. Increasing the size of the context windows reduced the performance of the NasTa-F

which achieved correlations below the average of the correlation of the human performance as shown in figure 6.8 and table 6.19. This result confirms the assumption that words closer to the target word are more likely to be related than those which are further from the target word, thus usage of a small context window may well result in fewer irrelevant words being used as well as being computationally more efficient.

The NasTa-F based on the modern POS classification achieved a best correlation (r = 0.901) among others. The NasTa-F is performing well at (r = 0.901) with the average value of the correlations of human participants (r = 0.892). Figure 6.9 shows the correlation between the NasTa-F and human ratings on the ASTSS-68 dataset. Furthermore, the performance of the NasTa-F was substantially better than the worst human (lower bound) performance at (r = 0.80).

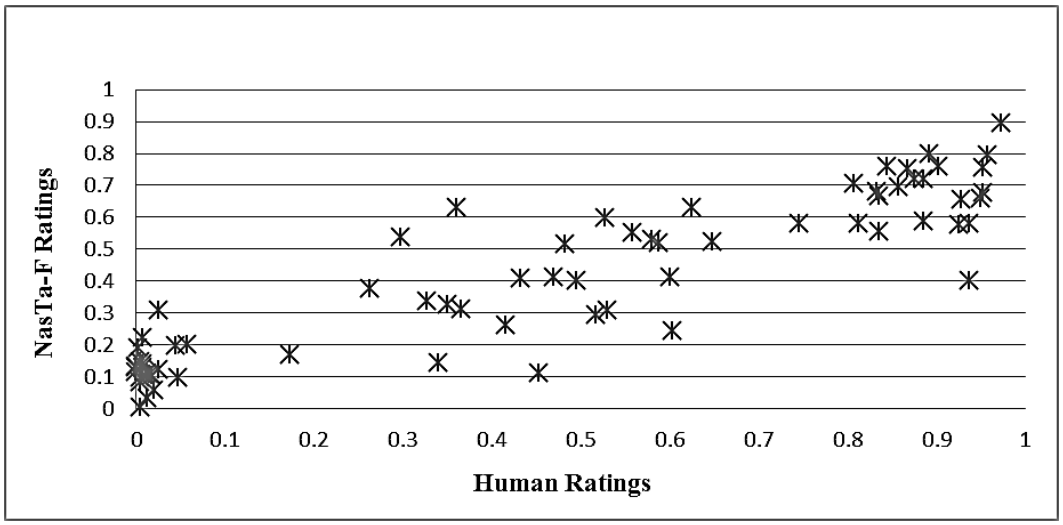


Figure 6.9 The Correlation between the Human Ratings and the NasTa-F without Word Order (window size 3).

6.3.3.3 Evaluation of NasTa-F with the Word Order

This section describes the evaluation process of the NasTa-F with the word order component to investigate its influence in the NasTa-F via comparing its performance with and without the word order component. In this case, the NasTa-F calculates the short text similarity based on the POS, WSD, semantic similarity and the word order similarity.

The evaluation methodology consisted of two steps. These are the determination of the optimal parameter values of the NasTa-F algorithm and the application of the ASTSS-68 dataset pairs on the NasTa-F algorithm with the word order similarity component. For the first step, NasTa-F requires determining the optimal values for four parameters before use. These are the window size for AWSAD, a threshold for the semantic vector derivation, a threshold for the word order vector formation and δ for adjusting the relative contributions of semantic and word order information to the final NasTa-F calculation.

In accordance with the same procedure used in optimising the NasTa-A algorithm parameters section (6.3.2.4), the ASTSS-21 set was used to determine the optimal parameter values for the NasTa-F algorithm. For the same consideration described in section (6.3.2.4), the initial values were given to each of the NasTa-F parameters whereby the initial value given to the semantic threshold parameter is 0.20, the word order threshold parameter is 0.30, the window size 3 and δ is 0.55.

Given the initial value of each parameter, the short text pairs on the ASTSS-21set were run using the NasTa-F with the word order to produce machine similarity ratings in a range of 0 to 1. The correlation coefficient between the human ratings of ASTSS-21set and those obtained from the NasTa-F was computed. The values of the Arabic algorithm parameters were changed to obtain a set of correlation coefficients. The increasing step for δ, semantic and word order thresholds parameters was 0.05 whilst the window size parameter was changed to 5 and 7.

The parameters with the strongest correlation coefficient were considered as the optimal parameters. In this experiment, the strongest correlation coefficient was obtained at $\delta = 0.55$, the semantic threshold = 0.2, the word order threshold = 0.70 and the window size =3.

Using the optimal parameter values, the short text pairs on the ASTSS-68 dataset were run using the NasTa-F algorithm in order to produce the machine similarity ratings in the range from 0 to 1.

The NasTa-F algorithm with the word order achieved a correlation (r = 0.876) below the correlation achieved by NasTa-F without the word order (r = 0.901) on the ASTSS-68 dataset. The results indicate that the presence of the word order similarity component has reduced the performance of the NasTa-F algorithm. Consequently the decision was made to remove the word order component from the NasTa-F algorithm and the short text similarity is calculated based POS, WSD and sematic similarity.

6.3.3.4 Comparison with the NasTa-A Performance

Steiger's z-test was used to compare the difference between the performance of the NasTa-F and NasTa-A algorithms on the ASTSS-68 dataset. Using Steiger's z-test requires the construction of a correlation triangle (3 correlations) between:

NasTa-A ratings vs. Human ratings = 0.785
NasTa-F ratings vs. Human ratings = 0.901
NasTa-A vs. NasTa-F = 0.841
n = 68 (the number of short text pairs in the ASTSS-68 dataset)

Applying the test (using the online calculator which was available at (Grabin, 2013)) indicates that the difference between NasTa-F and NasTa-A is statistically significant ($Z = -3.52$, $p < .001$).

This result indicates that extension of the NasTa-A algorithm for understanding context within a short text structure (by performing the Arabic WSD) and the use of POS other than nouns improved the algorithm performance. The NasTa-F algorithm has succeeded in obtaining similarity values close to human ratings for many short text pairs in each similarity range that the NasTa-A algorithm failed to obtain, as shown in figure 6.10 and table 6.20. Figure 6.10 shows the difference between the correlations achieved by the NasTa-A and the NasTa-F on the ASTSS-68 dataset. Table 6.20 presents the human similarity ratings with the corresponding machine similarity ratings produced by NasTa-A and NasTa-F algorithms on the ASTSS-68 dataset. As shown in figure 6.10 and table 6.20, the short text pairs rated high by participants such as (1, 14, 16, 21, 22, 28, 29, 50, 63, and 64) obtained low similarity or low medium similarity values by the NasTa-A whilst the NasTa-F improved the

similarity score and gave values close to human ratings. Also the NasTa-F obtained values closer to human ratings from many pairs that rated low or medium by participants such as (67, 60, 59, 55, 48, 41, 39, 31, 27, 24, 17 and 10) whilst the NasTa-A failed.

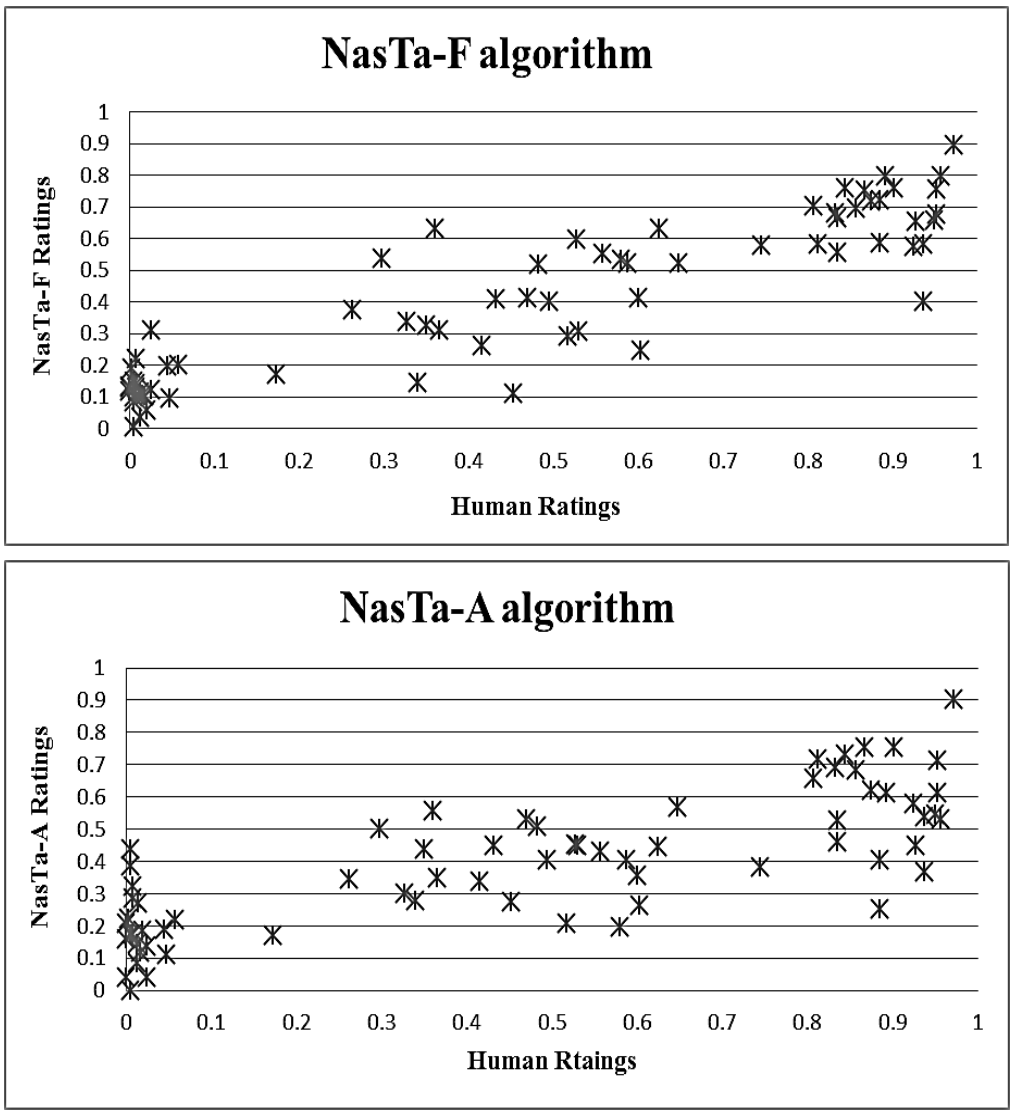


Figure 6.10 the correlations achieved by NasTa-A algorithm and NasTa-F algorithm on ASTSS-68 dataset.

Table 6.20 Short Text Similarity Ratings for ASTSS-68 dataset from Human, NasTa-A and NasTa-F.

ST Pairs	Human Ratings	NasTa-A Ratings	NasTa-F Ratings	ST Pairs	Human Ratings	NasTa-A Ratings	NasTa-F Ratings
1	0.95	0.54	0.66	35	0.97	0.90	0.90
2	0.85	0.73	0.76	36	0.93	0.58	0.58
3	0.53	0.45	0.31	37	0.60	0.26	0.24
4	0.42	0.34	0.26	38	0.47	0.53	0.41
5	0.00	0.22	0.19	39	0.58	0.20	0.53
6	0.02	0.19	0.06	40	0.05	0.19	0.20
7	0.86	0.68	0.69	41	0.01	0.32	0.22
8	0.90	0.75	0.76	42	0.87	0.75	0.75
9	0.95	0.61	0.68	43	0.95	0.71	0.75
10	0.60	0.35	0.41	44	0.65	0.57	0.52
11	0.33	0.30	0.34	45	0.37	0.35	0.31
12	0.01	0.14	0.11	46	0.00	0.04	0.12
13	0.06	0.22	0.20	47	0.26	0.34	0.37
14	0.96	0.53	0.80	48	0.02	0.27	0.11
15	0.83	0.69	0.68	49	0.94	0.54	0.58
16	0.89	0.25	0.72	50	0.89	0.61	0.80
17	0.59	0.40	0.52	51	0.48	0.51	0.52
18	0.36	0.56	0.63	52	0.50	0.40	0.40
19	0.02	0.12	0.10	53	0.01	0.00	0.08
20	0.01	0.08	0.03	54	0.03	0.14	0.31
21	0.75	0.38	0.58	55	0.01	0.38	0.01
22	0.81	0.66	0.70	56	0.81	0.72	0.58
23	0.53	0.45	0.60	57	0.94	0.37	0.40
24	0.63	0.44	0.63	58	0.34	0.28	0.14
25	0.43	0.45	0.41	59	0.35	0.44	0.33
26	0.00	0.16	0.13	60	0.00	0.21	0.13
27	0.01	0.28	0.14	61	0.01	0.17	0.15
28	0.84	0.53	0.67	62	0.17	0.17	0.17
29	0.88	0.62	0.72	63	0.89	0.40	0.59
30	0.84	0.46	0.56	64	0.93	0.45	0.66
31	0.56	0.43	0.55	65	0.52	0.21	0.29
32	0.45	0.27	0.11	66	0.30	0.50	0.54
33	0.05	0.11	0.10	67	0.01	0.44	0.10
34	0.01	0.17	0.13	68	0.03	0.04	0.12

Table 6.21 shows the difference between the performances of NasTa algorithms by means of comparison their performance with the average of the correlation of human participants. Whereby the performance of the NasTa-A at (r =0.785) was significantly below the average of human performance at (r = 0.892) and also below

the worst participants at (r = 0.80). Whilst the NasTa-F was performing well at (r = 0.901) with the average of human and also it was substantially better than the worst participants.

Table 6.21 The Performance of NasTa-A and NasTa-F Algorithms on the ASTSS-68 Dataset

On the ASTSS-68 Dataset	Correlation r
NasTa-A algorithm	0.785
NasTa-F algorithm	0.901
Average of the correlation of all participants	0.892
Best participants	0.970
Worst participants	0.80

Returning to the example in section 6.3.1.1 of the short text pair number 67 selected from the ASTSS-68 dataset (for the purpose of comparison with the NasTa-A performance), the NasTa-F calculates the short text similarity as follows:

T_1: انصحك ان تذهب إلى المكتبة وتقض فيها وقتاً نافعاً بين عقول المؤلفين

I recommend you to go to the library and to spend a quality time among thoughts of authors.

T_2: الذهب هو احد المعادن المهمة في عالم الاقتصاد ومحرك اساسي لحركة الاسواق

Gold is one of the important metals in the economic world and has an essential impact on the market movement.

Step1 includes assigning the POS to every word in the input short texts and determining the lemma for each word in the two short texts.

Step 2 involved disambiguating each word (nouns and verbs) in the two short texts. Each word is paired with the correct sense assigned to this word by the AWSAD algorithm.

Step 3 is to create the joint word set T for the short text T1 and the short text T2:

انصحك, ان, تذهب, الى, المكتبة,و تقضي, فيها, وقتا, نافعا, بين, عقول, المؤلفين, الذهب, هو, احد, المعادن, }
{المهمة, في, عالم, الاقتصاد, ومحرك, اساسي, لحركة, الاسواق

Step 4 involves the calculation of the semantic similarity component (S_S), where the semantic vectors for the two short texts T_1 and T_2 can be created from the joint word set T and corpus statistics. Table 6.22 illustrates the process of the creation of the semantic vector for T_1. The rightmost column in table 6.22 lists words in T whilst the first row lists words in the short text T_1. The words in the first column and row are listed in the order as they occur in the joint word set T and the short text T_1.

Table 6.22 The Semantic Vector Creation Process

		T_1 (w_i)												
Weight $I(w_i).I(\hat{w}_i)$	S	12.المؤلفين authors	11. عقول thoughts of	10.بين among	9. نافعا a quality	8. وقتا time	7. فيها in it	6. وتقض to spend	5.المكتبه the library	4. الى to	3. تذهب go	2. ان to	1. انصحك I recommend you	Joint word set $(\hat{w}_i)$
0.260	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	1. انصحك I recommend you
0.025	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	2. ان to
0.489	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	3. تذهب go
0.078	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	4. الى to
0.227	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	5. المكتبة the library
0.394	1.000	0.00	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	6. وتقض to spend
0.025	1.000	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	7. فيها in it
0.376	1.000	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8. وقتا time
0.321	1.000	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9. نافعا a quality
0.074	1.000	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10. بين among
0.197	1.000	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11. عقول thoughts of
0.212	1.000	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12. المؤلفين authors
0.000	**0.000**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	**13. الذهب gold**
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	14. هو is
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	15. احد one of
0.000	**0.000**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	**16. المعادن metals**
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17. المهمه the important
0.025	1.000	0.00	0.00	0.00	0.00	0.00	1.000	0.00	0.00	0.00	0.00	0.00	0.00	18. في in
0.000	**0.00**	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**19. عالم world**
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20. الاقتصاد economic
0.404	**0.215**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.215**	0.00	0.00	0.00	0.00	21. ومحرك impact
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22. اساسي an essential
0.000	**0.000**	0.00	0.00	0.00	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	**0.000**	23. لحركة movement
0.000	0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	24. الاسواق the markets

For each word in the joint word set T, if the same word occurs in the short text T_1 and they have the same POS then set the cross point cell to 1. For example, the word انصحك (I recommended you) appeared in T and T_1 (same form and POS) and the cell is set to 1, as shown in table 6.22.

For each word in the joint word set T, if the word has the same lemma and POS with any word in the short text T_1 then set the cross point cell to 1. For example, the word number 18 in the T في "in" and the word number 7 in T_1 فيها "in it", they have the same lemma في and the same POS (preposition). Thus the cell at the cross point is set to 1.

If the word has the same lemma with any word in the short text T_1 but different POS then set the cross point cell to 0. For example, the noun الذهب "gold" and the verb تذهب "go", they have the same lemma ذهب but different POS (noun "gold" and verb "go") so the cell at the cross point is set to 0. These two words were considered as the same word in the NasTa-A algorithm and the cell at the cross point was set to 1.

Otherwise, the cell at the cross point of the most similar word (same POS either pair of nouns or verbs) is set to their similarity value or 0 (different POS), reliant on whether the highest similarity score exceeds the threshold of 0.2. The KalTa-A and KalTa-F measures calculated the similarity between two nouns or verbs using the correct sense assigned to each word in T and T_1 by the AWSAD algorithm.

For example, the word number 19 in T عالم offers multiple meanings which mean عالَم ***Aalam*** "world" or عالِم ***Aalim*** "scientist". The correct sense assigned to this word by AWSAD is عالَم ***Aalam*** "world" and there is no similarity between this word and any word in the T_1 so is set to 0. The NasTa-A algorithm calculated the similarity without WSD and considered the word عالم as "scientist" not "world" and gave a high medium similarity with the word authors. For the same reason (using the correct sense), the cell at the cross point of the word لحركة "movement" and وقتا "time" is set to 0 whilst the NasTa-A gave a medium similarity value between them.

Also the NasTa-A considered the word تذهب "go" as a noun (gold) and gave a high medium similarity with the word metals whilst the NasTa-F considered them have different POS and the cell is set to 0. For the same reason, the (different POS) the cell at the cross point of the word لحركة "movement" and انصحك "I recommended you" is set to 0 whilst the NasTa-A gave a medium similarity value between them.

Step 5: The largest value in each row is chosen to create the lexical vector s_1 for the T_1. The leftmost column lists the corresponding information content to weight the significance of the word. Consequently, the semantic vector for the short text T_1 is:

s_1: {0.260, 0.025, 0.489, 0.078, 0.227, 0.394, 0.025, 0.376, 0.321, 0.074, 0.197, 0.212, 0.0, 0.0, 0.0, 0.0, 0.0, 0.025, 0.0, 0.0, 0.087, 0.0, 0.0, 0.0}.

In accordance with the same process, the semantic vector for the short text T_2 is:
s_2: {0.0, 0.0, 0.0, 0.0, 0.087, 0.0, 0.025, 0.0, 0.0, 0.0, 0.0, 0.0, 0.179, 0.081, 0.144, 0.363, 0.154, 0.025, 0.107, 0.149, 0.716, 0.177, 0.113, 0.138}.

Using s_1 and s_2, the semantic similarity between T_1 and T_2 is $S_S = 0.10$.

The NasTa-F gave this short text pair (number 67) a very low similarity value (0.10) which was very closer to the human assessment (0.02). Whilst the NasTa-A gave this pair of short text a medium similarity value (0.44) as shown in table 6.20.

This result indicates that the calculation of the Arabic short text semantic similarity based on the POS and AWSD improved the performance of the NasTa and this combination should be used profitably in the NasTa framework.

6.4 Conclusions

This chapter has described the production of the first Arabic short text benchmark dataset (ASTSS-68) with its creation methodology. The motivation of the creation of this dataset was to evaluate the Arabic short text similarity framework (NasTa) presented in chapter 4. It is expected that ASTSS-68 will make a substantial contribution to future work in the field of Arabic short text semantic similarity and hopefully it will be considered as a reference basis from which to evaluate and compare different methodologies in the field.

The creation methodology involved two experiments: the first was to produce the materials and the second was to collect human ratings. The experiment to create the

materials included selecting the set of 68 stimulus words covered a range of Arabic language features by means of populating a sampling frame, generating a database of 1088 Arabic short texts using a sample of 32 native Arabic speakers with a capacity for creative writing and finally, selecting the set of 68 short text pairs from the database which covered a varying range of similarity. This was followed by selecting a set of short text pairs nominated by three judges from the database which pilot ratings by a small sample of human participants in order to select the final set with greater confidence before running the rating experiment.

Human ratings were collected for 68 short text pairs in accordance with the same procedure used to collect human ratings in Noun and Verb datasets (chapter 5). The sample of participants used in this experiment was selected to achieve a balance and also representation of the human population. Good care was taken to control the distribution of the participants' age, academic background, educational level and gender. The results of this experiment were reported using Pearson correlation coefficients. The average of the correlation of all participants was calculated and this can be used to assess the performance of a computational method attempting to perform the same task.

This chapter also described the evaluation procedure of NasTa which consisted of four major steps:

Step 1 included creation of an optimization dataset in order to determine the optimal parameter values of the NasTa. A set of 21 short text pairs (ASTSS-21) with human ratings covering a varying range of similarity was created using the rest of the short text pairs nominated by three judges in ASTSS-68 dataset.

Step 2 included evaluation of the NasTa-A algorithm as created in the first phase of the NasTa framework development process. The evaluation methodology included determination of the optimal parameter values of the NasTa-A algorithm using the ASTSS-21 dataset and the application of the ASTSS-68 dataset pairs on the NasTa-A algorithm. The optimal value of the semantic threshold parameter was 0.2 and the δ parameter was 1 whilst the optimal parameter values used with Arabic noun (KalTa-A) measure were $\alpha = 0.12$ and $\beta = 0.21$. The results of the evaluation process

indicated that the NasTa-A algorithm performed significantly below the average human performance. Two issues affected the performance of NasTa-A. It focused on the similarity of nouns only and did not take the context in which the nouns occur into account. Additionally, the evaluation methodology included investigation of the influence of word order similarity in the NasTa-A algorithm. The result from Steiger's z-test indicated that the word order similarity had no influence on the performance of the NasTa-A algorithm.

Step 3 included the evaluation of the NasTa-F algorithm which was evaluated in accordance with the same procedure used to evaluate the NasTa-A algorithm. The optimal parameter values of NasTa-F were determined. The semantic threshold was 0.2, the word order threshold was 0.70 and the δ parameter was 0.55 whilst the optimal parameter values used with Arabic verb (KalTa-F) measure were $\alpha = 0.2$ and $\beta = 0.459$. The window sizes tested with WSD algorithm were 3, 5 and 7. The performance of NasTa-F algorithm with and without the word order was also investigated. The NasTa-F with window size 3 and without the word order component achieved a best correlation performing well compared with the average human performance. The presence of the word order component reduced the performance of the NasTa-F algorithm. Consequently, the decision was made to remove the word order component from the NasTa-F algorithm.

Step 4 involved the determination of which combination should be used profitably in NasTa framework. Steiger's z-test was used for this purpose and the results indicated that the NasTa-F algorithm performed better than the NasTa-A algorithm. The improvement achieved was statistically significant at $P < 0.001$. The results also indicated that the ratings from the computational short text semantic similarity can be improved by means of understanding context within a short text structure and the use of POS other than nouns.

Chapter 7

Conclusions and Future Work

7.1 Summary of Contributions

The contribution of the work in this thesis falls into three areas: Arabic semantic similarity measures, Arabic Word Sense Disambiguation (WSD) and Arabic semantic similarity resources. Figure 7.1 presents the contributions of this work in each area.

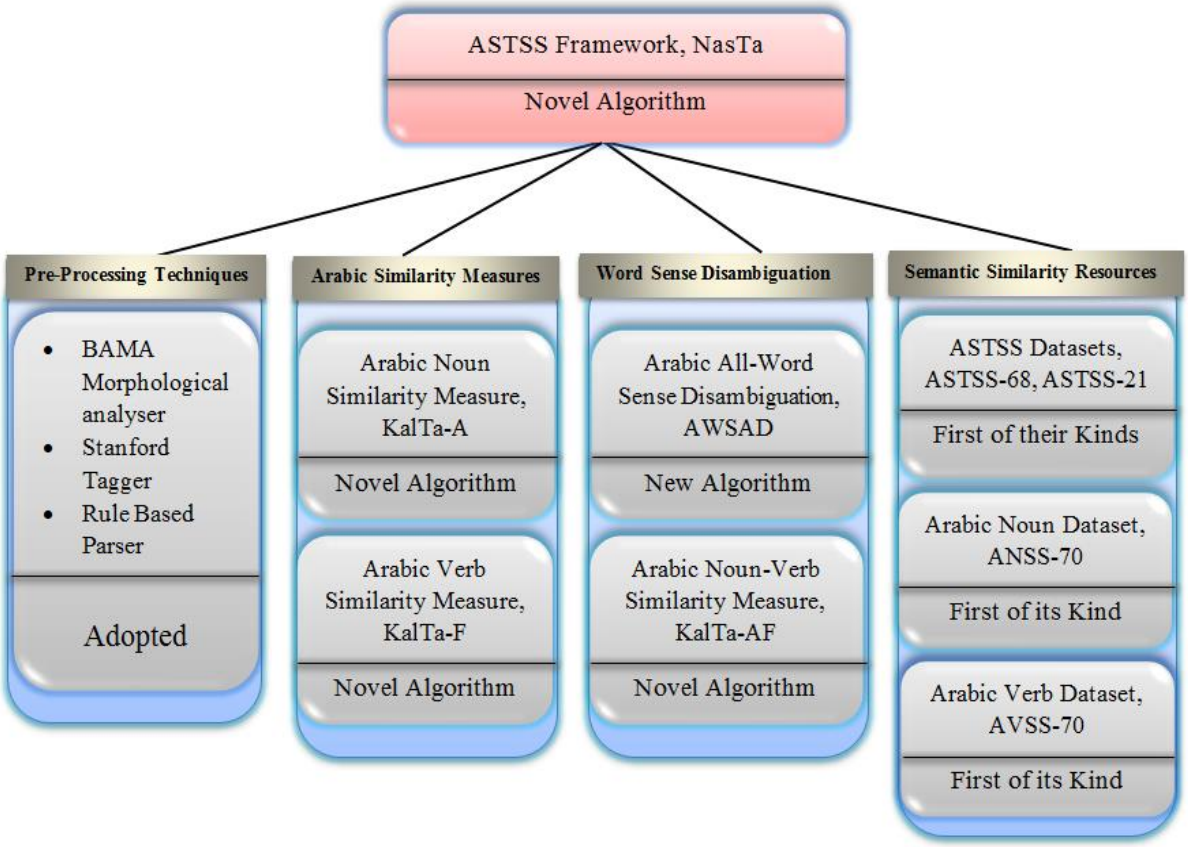


Figure 7.1 The Contributions of this Work in different areas.

As shown in figure 7.1, the main contribution of the work in this thesis is the presentation of a novel framework (NasTa) for developing an Arabic Short Text Semantic Similarity (STSS) measure which calculates the similarity between two short texts based on POS, Arabic WSD and semantic similarity. The modularity of

the framework allows new or improved components to be incorporated in framework in future. This contribution falls into the Arabic semantic similarity measure area. Many Arabic applications can benefit from the use of an Arabic STSS measure such as Conversation Agents, Text Mining and Information Retrieval. Other original contributions include:

1. Arabic Semantic Similarity Measures

- A novel Arabic Noun Semantic Similarity measure (KalTa-A) to identify the similarity score between two Arabic nouns.
- A novel Arabic Verb Semantic Similarity measure (KalTa-F) to calculate the similarity between pairs of Arabic verbs.
- A novel algorithm (KalTa-AF) was presented to identify the similarity score between two words which had a different POS, either a pair comprising a noun and verb or a verb and noun. This algorithm was developed to perform Arabic WSD based on the concept of noun semantic similarity.

2. Arabic Semantic Similarity Resources

- The production of the first Arabic noun benchmark dataset (ANSS-70) for evaluating noun similarity algorithms. Moreover, two sub-datasets known as training and evaluation were specified by partitioning the noun benchmark which can be used for training and testing different methodologies.
- The production of the first Arabic verb benchmark dataset (AVSS-70) for the evaluation of the verb similarity algorithms. Training and evaluation sub-datasets were specified to train and test different verb methodologies.
- The production of the first Arabic short text benchmark dataset (ASTSS-68) for evaluating STSS measures. An optimization dataset (ASTSS-21) was also produced which can be used in tuning or optimizing the algorithms.

These datasets will make a substantial contribution to future research in the field of Arabic word and short text semantic similarity. Specification of the partition supports objective comparison of new trainable algorithms as the field develops. It is to be hoped that this will be regarded as a reference basis from which to evaluate and

compare different methodologies in the field. Furthermore, the procedures used for production of these datasets can be used by other Arabic researchers to extend them.

3. Arabic word sense disambiguation

A new algorithm for Arabic WSD namely that of AWSAD was presented to disambiguate all words (nouns and verbs) in the Arabic short texts based on a knowledge-based approach. The AWSAD algorithm performed WSD without requiring any manual training data whilst used the AWN as a knowledge base.

7.2 Summary of Work

This thesis has presented a novel framework for the development of a STSS measure for Modern Standard Arabic (MSA) and implemented a measure within that framework. At the onset of the work, a search of the literature showed that no STSS measure had been undertaken for MSA and this required investigation in three directions as follows:

1. The characteristics of the Arabic language and their influence on STSS computation. This step considered the research question 'Are there features of Arabic language which would prevent the construction of the framework for semantic similarity?' A thorough review of the literature in chapter 2 has shown that the complex internal word structure, missing diacritics and syntactical flexibility features posed interesting challenges to the Arabic STSS computation.

2. The STSS creation requirements and determination of the drawbacks of the current state of the STSS measures. The research question 'Do the necessary components exist for constructing a measure with a framework?' was considered in this step. The majority of current short text measures rely largely on methods for composing an STSS measure from word similarity measures. A search of the literature showed that no word semantic similarity measure had been undertaken for MSA. Moreover, the literature search demonstrated that the major challenges faced by existing STSS measures consisted of understanding the context within a short text structure and the use of POS over and above nouns.

3. The methodologies used for the evaluation of the STSS measure. The only way to identify the quality of a machine STSS measure was by means of the use of a benchmark dataset with similarity ratings collected from human participants. No STSS benchmark dataset had been reported in the literature for MSA. The research question 'Is it possible to create suitable benchmark dataset for STSS algorithm test?' was considered.

Chapter 4 presented an investigation into the main research question 'Is it possible to construct a framework for developing a short text semantic similarity measure for Arabic language?' The investigation process comprised two phases.

First phase concerned the creation of an Arabic STSS algorithm, namely that of NasTa-A which was inspired by Li et al.'s algorithm. The NasTa-A algorithm consisted of two fundamental components: the semantic similarity component and the word order similarity component. The computation process of the two components relied on the computation of the noun semantic similarity in both short texts. As mentioned earlier, no word (noun) semantic similarity measure had been undertaken for MSA. The research question 'Where there are missing components from NLP that are required, is it possible to create these for the Arabic language?' was considered in this phase. A new algorithm (KalTa-A) was created to identify the similarity between pairs of Arabic nouns using a knowledge based approach requiring information sources extracted from the lexical database AWN and taking advantage of the mapping with SUMO. As a consequence of the nature of the AWN organization scheme, the structure of its hierarchy may produce a bias towards a particular distance computation. The KalTa-A measure was hampered by this weakness as its recall relies on the AWN ontological detail and coverage. AWN was mapped to the SUMO ontology and the KalTa-A measure took advantage of this mapping to overcome its limitations.

The research question 'Is it possible to create suitable benchmark dataset for noun algorithm test?' was investigated in the first phase. This thesis described the creation of the first Arabic noun (ANSS-70) dataset and its production methodology which involved two experiments: the first was to produce the materials and the second was to collect human ratings. This research used a systematic process in the creation materials experiment whereby a new method was used to select the set of 56 stimulus

nouns by means of the creation of 27 Arabic categories with 27 different themes to promote the best possible semantic representation. Unlike the prior work in English word similarity, 22 participants were chosen to produce a set of 70 noun pairs which covered a range of semantic similarity values from maximum to minimum. Human ratings were collected from a new sample of 60 participants using the best possible available techniques. Care was taken to control the distribution of the participants' age, academic background, educational level and gender. Based on review of prior work in English and consistency between participants, the current evidence supports rejecting the null (research) hypothesis 'it is not possible to construct a noun dataset for Arabic within a limited size which effectively represents human intuition'.

Training and evaluation datasets were produced using the ANSS-70 dataset in order to apply them in the evaluation procedure of the Arabic noun measure. The training dataset was used in the optimization of parameters in the algorithm whilst the quality of the noun measure was identified using the evaluation dataset. Experimental evaluation indicated that the noun measure achieved a good correlation at $r = 0.91$ compared with the average human performance at $r = 0.893$. Since the results from the Arabic noun algorithm exceed the average human performance ($r = 0.893$), it will be considered performing well and the null (research) hypothesis 'it is not possible for a machine based Arabic noun semantic similarity measure to re-produce human intuitive measures of semantic similarity.' can be rejected. These results also answered the research question 'Is it possible to measure the semantic similarity between a pair of Arabic nouns?' The Arabic noun measure with its optimal parameter values ($\alpha = 0.12$ and $\beta = 0.21$) was used with the NasTa-A short text algorithm.

The computation of the semantic similarity component utilised information extracted from a structured lexical database AWN and corpus statistics known as the Arabic Word Count (AWC). The BAMA morphological analyser and the Stanford POS tagger were selected based on their accuracy and availability to address the challenge of the complex internal structure of Arabic words which prevents the extraction of semantic information from AWN and AWC directly. NasTa-A incorporated syntactic information by forming the word order vector for each short text based on a word sequence and location in a short text. Attai's Rule Based parser was adopted within

this framework to address the syntactical flexibility of MSA to take advantage of the word order which contributed to the Li measure. The research question 'Do the necessary components exist for constructing a measure with a framework?' was considered.

The next step of the work was to evaluate the NasTa-A algorithm. The research question 'Is it possible to create suitable benchmark dataset for STSS algorithm test?' was investigated in this step. The first Arabic short text benchmark dataset (ASTSS-68) was created. The creation materials experiment included selecting a set of 68 stimulus words which covered a range of Arabic language features via populating a sampling frame, generating a database of 1088 Arabic short texts using a sample of 32 native Arabic speakers with a capacity for creative writing, and finally, selecting the set of 68 short text pairs from the database which covered a varying range of similarity based on human judgements. Human ratings were collected from a new sample of 62 participants in accordance with the same procedure used to collect human ratings in the noun dataset. This dataset took a good care to control the distribution of the participants' age, academic background, educational level and gender. Based on review of prior work in English and consistency between participants, the current evidence supports rejecting the null hypothesis 'it is not possible to construct a short text dataset for Arabic within a limited size which effectively represents human intuition'.

An optimization dataset of 21 short text pairs (ASTSS-21) was created using the remainder of the short text pairs nominated by three judges and rated by 10 participants in the ASTSS-68 dataset. The ASTSS-21 was used to determine the optimal parameter values of the NasTa-A algorithm whilst the ASTSS-68 dataset was used to identify the quality of the NasTa-A algorithm. Experimental evaluation indicated that the NasTa-A at $r = 0.785$ performed significantly below the average human performance at $r = 0.892$ and the word order similarity component had no influence on the performance of the NasTa-A algorithm. At this stage, it was not possible to reject the null hypothesis 'it is not possible for a machine based Arabic STSS measure to re-produce human intuitive measures of semantic similarity'. The unexpected performance of the NasTa-A resulted from the missing diacritics feature of MSA and the drawbacks of the existing STSS measures which focused only on the

similarity of nouns and did not take the context in which the nouns occur into account.

Further research was required to extend the NasTa-A algorithm in order to improve its performance by means of understanding the context within a short text structure and the use of POS other than nouns. Consequently, the second phase of the development process of the NasTa framework involved developing a new ASTSS algorithm, NasTa-F, which covered the POS, Arabic WSD, semantic similarity and word order similarity.

The computation process of semantic and word order components were based on the POS which used the noun measure (KalTa-A) to calculate the similarity between pairs of nouns. Adjective and adverb pairs either had exact lexical matches, whereby both came from the same POS or were rated as unrelated in meaning. Finally a novel algorithm (KalTa-F) was presented to calculate the similarity between pairs of verbs based on the assumption that words sharing a common root usually have a related meaning.

The research question 'Is it possible to create suitable benchmark dataset for verb algorithm test?' was investigated in this phase. This thesis described the production of the first Arabic verb dataset (AVSS-70) and its creation methodology. In the creation materials experiment, a set of 25 stimulus verbs was selected by decomposing the Arabic verbs into a tree structure based on special syntactical and semantic features. Unlike previous research studies, participants were chosen to produce a set of 70 verb pairs which covered a range of semantic similarity values from maximum to minimum. Human ratings were collected from a new sample of 60 participants in accordance with the same procedure used to collect human ratings in noun dataset. Care was taken to control the distribution of the participants' age, academic background, educational level and gender. Based on review of prior work in English and consistency between participants, the current evidence supports rejecting the null hypothesis 'it is not possible to construct a verb dataset for Arabic within a limited size which effectively represents human intuition'.

Training and evaluation datasets were produced using the AVSS-70 dataset in order to apply them in the evaluation procedure of the Arabic verb measure. Experimental

evaluation indicated that the verb measure performed well and achieved a good correlation at $r = 0.906$ which exceeded the average human performance at $r = 0.887$. The null hypothesis 'it is not possible for a machine based Arabic verb semantic similarity measure to re-produce human intuitive measures of semantic similarity' was rejected. This result answered the research question 'Is it possible to measure the semantic similarity between a pair of Arabic verbs?' and the Arabic verb measure with its optimal parameters values was used with the NasTa-F algorithm.

The second phase also considered the research question 'Is it possible to disambiguate all words in an Arabic short text?' by presenting a new knowledge-based Arabic WSD algorithm (AWSAD) to disambiguate all words (nouns and verbs) in the Arabic short texts which relied on AWN similarity measures including the noun measure and verb measure. A novel measure was presented to identify the similarity between two words which had a different POS, either a pair comprising a noun and verb or vice-versa. This measure was created to overcome the limitations of AWSAD algorithm and the research question 'Is it possible to measure the similarity between Arabic words belonging to a different POS?' was considered. The AWSAD algorithm was employed by NasTa-F to address the challenge of missing diacritics in contemporary Arabic writing causing great ambiguity. No Arabic manually sense-tagged data was available to evaluate the AWSAD algorithm therefore it was decided to evaluate this algorithm in terms of its performance within the NasTa-F algorithm. This was feasible because every other component of NasTa-F had been evaluated in isolation as well as within NasTa-F, avoiding confounding factors.

The NasTa-F algorithm was evaluated in accordance with the same procedure used to evaluate the NasTa-A algorithm. The optimal parameter values of NasTa-F were determined using the ASTSS-21 dataset. The performance of the NasTa-F algorithm, with and without the word order, was investigated. The NasTa-F with window size 3 and without the word order component achieved the best correlation which performed well at $r = 0.901$ with the average human performance at $r = 0.892$. The null hypothesis 'it is not possible for a machine based Arabic STSS measure to re-produce human intuitive measures of semantic similarity' was rejected and the main

research question ‘Is it possible to construct a framework for developing a short text semantic similarity measure for Arabic language?’ was answered.

The presence of the word order component reduced the performance of the NasTa-F algorithm at $r = 876$ which resulted from the feature of the complex internal structure of the Arabic words. Consequently, the decision was made to remove the word order component from the NasTa-F algorithm. Finally, experimental results indicated that the NasTa-F algorithm performed ($r = 0.901$) significantly better than the NasTa-A algorithm ($r = 785$) and indicated that the ratings from the computational STSS could be improved by means of understanding context within a short text structure (using WSD) and the use of POS over and above nouns. Based on this result the null hypothesis ‘it is not possible for an Arabic algorithm for all-word sense disambiguation to achieve the same classification as human would make’ was rejected and the research question ‘Is it possible to disambiguate all words in an Arabic short text?’ was answered.

In summary, in each case evidence was found to reject the null hypothesis from the derived pairs of hypotheses. These conclusions are, of course, pending replication of results by independent researches joining this new and exciting field in the future.

7.3 Further Research

This section will focus on the NasTa framework components that would take advantage of further research.

7.3.1 Semantic Similarity

Although the improvement achieved by means of understanding context within a short text structure and the use of POS was statistically significant, the NasTa-F algorithm has a limitation. The computation process of a semantic similarity component involved comparing pairs of words belonging to the same POS. These pairs comprised either pairs of nouns or pairs of verbs. Whilst the adjective and adverb pairs either had exact lexical matches whereby both came from the same POS or were rated as unrelated in meaning. Oliva et al. (2011) provided evidence that

adjectives and adverbs play an important role in short text semantics and should be used in the short text similarity computation. Consequently, further research is required in order to involve similarities of adjectives, adverbs and words belonging to a different POS in the computation process of semantic similarity component.

In order to compare a pair of noun and verb within the short text similarity component, the noun-verb similarity algorithm developed in this thesis to perform WSD can be used for this purpose with some modification to utilise a pair of words (noun and verb) instead of a pair of senses.

The gloss-based measure considers a suitable method (the only method) to calculate the similarities of adjectives and adverbs (Oliva et al., 2011 and Gou and Diab, 2009). However, Oliva et al. (2011) provided evidence that this measure is not appropriate for use with short text similarity computation. The gloss-based measure calculates the similarity score based on the overlap of the glosses associated with the concepts containing the compared words. Further research is required to develop this measure in order that it is suitable for calculating the adjective and adverb similarities within the short text similarity computation. For example, instead of using the overlap of the glosses, it should be possible to investigate the use of the nouns of each gloss to calculate the similarities of adjectives and adverbs by means of the calculation of the shortest path length and the depth of the compared nouns. This method can also be used to compare words belonging to different POS such as a pair of noun and adjective, a pair of verb and adjective, etc. Finally, the WSD can benefit from this measure to improve the Arabic WSD algorithm performance.

7.3.2 Arabic Word Sense Disambiguation

Chapter 4 presented a new Arabic WSD algorithm which was created to disambiguate all the words in the Arabic short text in order to improve the NasTa performance. As described in chapter 2, the evaluation process of the WSD algorithm's performance required Arabic manually sense-tagged data whose creation methodology needed human experts to be very familiar with each Arabic word's

definition. Diab et al. (2007) presented an Arabic all-words sense annotated set in running text but it is not available from the authors for the purpose of research.

Accordingly, there is a need to go beyond the simple evaluation of Arabic WSD in chapter 5 for the creation of a set of all- words sense annotated data for Arabic to be used in the all-words WSD evaluation process and also for it to be a freely available to Arabic researchers.

7.3.3 Arabic benchmark datasets

The word benchmark datasets, noun and verb with their sub-datasets (training and evaluation) require to be expanded by means of generation of more word (noun or verb) pairs with human similarity ratings. This will support more extensive testing, training, tuning and optimizing different methodologies and will provide more credible comparisons use new STSS algorithms emerge. The procedures used for production of these datasets can be used to expand them. This requires expanding the set of stimulus words (nouns or verbs) by means of:

- Using the additional Arabic categories created in ASTSS-68 dataset for noun.
- Using more Arabic VerbNet classes with Case Grammar's frames.

The short text dataset ASTSS-68 and the optimization dataset ASTSS-21 also require to be expanded. First, more short text pairs that support the similarity across the POS can be added to both datasets and this will support the validation of the STSS measures which use the word similarity measures that cross POS in the STSS computation. For example, the noun زيارة Z***iyara*** "visit" and the verb زار ***Zara*** "visit" can be used to create pairs of high, medium and low short texts by means of asking participants to write short texts using these words with a specific theme.

Second, expanding the size of the two datasets especially the small dataset ASTSS-21 to reinforce more extensive testing, training, tuning and optimizing different STSS methodologies. Larger volumes of data will support more diverse and robust machine learning techniques. This can be undertaken in accordance with the same

procedure used in creation of these datasets which requires increasing the number of the Arabic stimulus words and themes.

To sum things up, this thesis has presented the first steps in a new field of Arabic short text semantic similarity. A viable framework was developed for such measures and a functioning algorithm created. However, each component of the framework offers scope for future research activity in the field and the framework itself may be adapted by other researchers.

References

ABDUL-RAOF, H. 2000. *Arabic stylistics: A coursebook*, Otto Harrassowitz Verlag.

ACHANANUPARP, P., HU, X. & SHEN, X. 2008. The evaluation of sentence similarity measures. *Data Warehousing and Knowledge Discovery.* Springer.

AGIRRE, E., DIAB, M., CER, D. & GONZALEZ-AGIRRE, A. Semeval-2012 task 6: A pilot on semantic textual similarity. Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, 2012. Association for Computational Linguistics, 385-393.

AHMED, F. & NURNBERGER, A. 2009. Corpora based Approach for Arabic/English Word Translation Disambiguation. *Speech and Language Technology,* 11**,** 195-214.

Al-AZAWI, S. & BURADAH, H. 1976. Qimat al-Kalimat al-ShAiah fi Kutub al-Atfal. Egypt: al-Hayah al-Misriyyah al-Amah li al-Kitab.

AL-QAHTANI, D. M. 2005. *Semantic valence of Arabic verbs*, Librairie du Liban Publishers.

AL-SHALABI, R. M., AL SERHAN, H. & KANNAN, G. New approach for extracting Arabic roots. The International Arab Conference on Information Technology (ACIT), Egypt, 2003.

AL-SHAMMARI, E. & LIN, J. A novel Arabic lemmatization algorithm. Proceedings of the second workshop on Analytics for noisy unstructured text data, 2008. ACM, 113-118.

AL-SULAITI, L. & ATWELL, E. S. 2006. The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics,* 11**,** 135-171.

AL AMEED, H., AL KETBI, S., AL KAABI, A., AL SHEBLI, K., AL SHAMSI, N., AL NUAIMI, N. H. & AL MUHAIRI, S. S. Arabic light stemmer: A new enhanced approach. The Second International Conference on Innovations in Information Technology (IIT'05), 2005.

ALKHALIFA, M. & RODRÍGUEZ, H. 2010. Automatically Extending Named Entities coverage of Arabic WordNet using Wikipedia. *International Journal on Information and Communication Technologies*.

ATKINSON-ABUTRIDY, J., MELLISH, C. & AITKEN, S. 2004. Combining information extraction with genetic algorithms for text mining. *Intelligent Systems, IEEE,* 19**,** 22-30.

ATTIA, M., PECINA, P., TORAL, A., TOUNSI, L. & VAN GENABITH, J. 2011. A lexical database for modern standard Arabic interoperable with a finite state morphological transducer. *Systems and Frameworks for Computational Morphology.* Springer.

ATTIA, M. A. 2008. *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation.* University of Manchester.

AYTAR, Y., SHAH, M. & LUO, J. Utilizing semantic word similarity measures for video retrieval. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008. IEEE, 1-8.

BAALBAKI, M. 1987. *Al-mawrid: a modern English-Arabic dictionary*, Dar el-ilm lil-Malayen.

BAALBAKI, M. 2005. *Al-mawrid: an Arabic dictionary,* Dar el-ilm lil-Malayen.

BAAYEN, R., PIEPENBROCK, R., & VAN , R. (1993). The *{CELEX}* lexical data base on *{CD-ROM}*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

BANERJEE, S. & PEDERSEN, T. Extended gloss overlaps as a measure of semantic relatedness. IJCAI, 2003. 805-810.

BATTIG, W. F. (1979) Citation Classic: Battig W F & Montague W E. Category norms for verbal items in 56 categories: a replication and extension of the Connecticut category norms. J. Exp. Psychol. Monograph Suppl. 80: No. 3, Part 2, 1969. IN GARFIELD, E. (Ed.).

BATTIG, W. F. & MONTAGUE, W. E. 1969. Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology,* 80**,** 1-46.

BEESLEY, K. R. Finite-state morphological analysis and generation of Arabic at Xerox Research: Status and plans in 2001. ACL Workshop on Arabic Language Processing: Status and Perspective, 2001. 1-8.

BELKRIDEM, F. & EL SEBAI, A. 2009. An ontology based formalism for the arabic language using verbs and derivatives. *Communications of the IBIMA,* 11**,** 44-52.

BIKEL, D. M. Design of a multi-lingual, parallel-processing statistical parsing engine. Proceedings of the second international conference on Human Language Technology Research, 2002. Morgan Kaufmann Publishers Inc., 178-182.

BLALOCK, H. M. JR.(1979), Social statistics. McGraw Hill, New York.

BRESNAN, J. 2001. *Lexical-functional syntax*, Wiley-Blackwell.

BROWN Corpus Information, http://www.essex.ac.uk/w3c/corpusling/content/corpora/list/private/brown/brown.html, 2005.

BUCKWALTER, T. 2002. Buckwalter {Arabic} Morphological Analyzer Version 1.0.

BUHRMESTER, M., KWANG, T. & GOSLING, S. D. 2011. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science,* 6**,** 3-5.

BUTT, M., DYVIK, H., KING, T. H., MASUICHI, H. & ROHRER, C. The parallel grammar project. Proceedings of the 2002 workshop on Grammar engineering and evaluation-Volume 15, 2002. Association for Computational Linguistics, 1-7.

CALLISON-BURCH, C., FORDYCE, C., KOEHN, P., MONZ, C. & SCHROEDER, J. (Meta-) evaluation of machine translation. Proceedings of the Second Workshop on Statistical Machine Translation, 2007. Association for Computational Linguistics, 136-158.

CALLISON-BURCH, C., FORDYCE, C., KOEHN, P., MONZ, C. & SCHROEDER, J. Further meta-evaluation of machine translation. Proceedings of the Third Workshop on Statistical Machine Translation, 2008. Association for Computational Linguistics, 70-106.

CARAMAZZA, A. & SHELTON, J. 1998. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Cognitive Neuroscience, Journal of,* 10**,** 1-34.

CHALABI, A. Sakhr Arabic Lexicon. NEMLAR International Conference on Arabic Language Resources and Tools, 2004. 21-24.

CHARLES, W. G. 2000. Contextual correlates of meaning. *Applied Psycholinguistics,* 21**,** 505-524.

COOK, W. A. 1979. *Case grammar: development of the matrix model (1970-1978)*, Georgetown University Press Washington, DC.

DAIMI, K. 2001. Identifying syntactic ambiguities in single-parse Arabic sentence. *Computers and the Humanities,* 35**,** 333-349.

DAVIES, J. & WEEKS, R. QuizRDF: Search technology for the semantic web. System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on, 2004. IEEE, 8 pp.

DE BONI, M. & MANANDHAR, S. The Use of Sentence Similarity as a Semantic Relevance Metric for Question Answering. New Directions in Question Answering, 2003. 138-144.

DIAB, M., ALKHALIFA, M., ELKATEB, S., FELLBAUM, C., MANSOURI, A. & PALMER, M. Semeval 2007 task 18: Arabic semantic labeling. Proceedings of the 4th International Workshop on Semantic Evaluations, 2007. Association for Computational Linguistics, 93-98.

DIAB, M., HACIOGLU, K. & JURAFSKY, D. Automatic tagging of Arabic text: From raw text to base phrase chunks. Proceedings of HLT-NAACL 2004: Short Papers, 2004. Association for Computational Linguistics, 149-152.

DIAB, M. T. An unsupervised approach for bootstrapping Arabic sense tagging. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, 2004. Association for Computational Linguistics, 43-50.

DICHY, J. & FARGHALY, A. Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built. The MT-Summit IX workshop on Machine Translation for Semitic Languages, New Orleans, 2003.

DICHY, J. & HASSOUN, M. Some aspects of the DIINARMBC research programme. In the 6th ICEMCO, Cambridge, England. 1998.

ELKATEB, S., BLACK, W., RODRÍGUEZ, H., ALKHALIFA, M., VOSSEN, P., PEASE, A. & FELLBAUM, C. Building a wordnet for arabic. Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006), 2006a.

ELKATEB, S., BLACK, W., VOSSEN, P., FARWELL, D., RODRÍGUEZ, H., PEASE, A. & ALKHALIFA, M. Arabic WordNet and the challenges of Arabic. Proceedings of Arabic NLP/MT Conference, London, UK, 2006b. Citeseer.

ELMOUGY, S., TAHER, H. & NOAMAN, H. Naïve Bayes classifier for Arabic word sense disambiguation. proceeding of the 6th International Conference on Informatics and Systems, 2008. 16-21.

FARGHALY, A. & SHAALAN, K. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP),* 8**,** 14.

FELLBAUM, C. 1998. *WordNet,* Wiley Online Library.

FENG, J., ZHOU, Y.-M. & MARTIN, T. Sentence similarity based on relevance. Proceedings of IPMU, 2008. 833.

FILLMORE, C. 1968. The case for case.

FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G. & RUPPIN, E. 2002. WordSimilarity-353 test collection.

FRAWLEY, W. 1992. Linguistic semantics. New Jersey: Lawrence Erlbaum Associates.

GRABIN, C. 2013. General statistics. http://psych.unl.edu/psycrs/statpage/regression.html.

GUO, W. & DIAB, M. T. Improvements to monolingual English word sense disambiguation. Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions, 2009. Association for Computational Linguistics, 64-69.

GUREVYCH, I. & NIEDERLICH, H. Computing semantic relatedness in german with revised information content metrics. Proceedings of" OntoLex, 2005. 28-33.

HABASH, N. Y. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies,* 3**,** 1-187.

HAJIC, J., SMRZ, O., BUCKWALTER, T. & JIN, H. Feature-based tagger of approximations of functional Arabic morphology. Proceedings of the Workshop on Treebanks and Linguistic Theories (TLT), Barcelona, Spain, 2005.

HIJJAWI, M., BANDAR, Z., CROCKETT, K. & MCLEAN, D. ArabChat: an Arabic Conversational Agent. Computer Science and Information Technology (CSIT), 2014 6th International Conference on, 2014. IEEE, 227-237.

HLIAOUTAKIS, A., VARELAS, G., VOUTSAKIS, E., PETRAKIS, E. G. & MILIOS, E. 2006. Information retrieval by semantic similarity. *International journal on semantic Web and information systems (IJSWIS),* 2**,** 55-73.

HO, C., MURAD, M. A. A., KADIR, R. A. & DORAISAMY, S. C. Word sense disambiguation-based sentence similarity. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010. Association for Computational Linguistics, 418-426.

HOVY, E., MARCUS, M., PALMER, M., RAMSHAW, L. & WEISCHEDEL, R. OntoNotes: the 90% solution. Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers, 2006. Association for Computational Linguistics, 57-60.

IBN AS-SARRAJ, A. M. 2009. Al-Usool fi an-Nahw. Ed. MuhammadUthman Cairo: Maktaba Ath-Thaqafa Ad-Diniya.

IDE, N. & VÉRONIS, J. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics,* 24**,** 2-40.

ISLAM, A. & INKPEN, D. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD),* 2**,** 10.

JACKENDOFF, R. 1983. *Semantics and cognition*, MIT press.

JIANG, J. J. & CONRATH, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of International Conference on Research in Computational Linguistics.

JURAFSKY, D. & MARTIN, J. H. 2000. *Speech & language processing*, Pearson Education India.

KAMIR, D., SOREQ, N. & NEEMAN, Y. A comprehensive NLP system for modern standard Arabic and modern hebrew. Proceedings of the ACL-02 workshop on Computational approaches to semitic languages, 2002. Association for Computational Linguistics, 1-9.

KAYE, A. S. 1972. Remarks on diglossia in Arabic: well-defined vs. ill-defined. *Linguistics,* 10**,** 32-48.

KELLER, F., GUNASEKHARAN, S., MAYO, N. & CORLEY, M. 2009. Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods,* 41**,** 1-12.

KENNEDY, A. & SZPAKOWICZ, S. Evaluating Roget's Thesauri. ACL, 2008. Citeseer, 416-424.

KHOJA, S. APT: Arabic part-of-speech tagger. Proceedings of the Student Workshop at NAACL, 2001. 20-25.

KHOJA, S. & GARSIDE, R. 1999. Stemming arabic text. *Lancaster, UK, Computing Department, Lancaster University*.

KIPPER, K., DANG, H. T. & PALMER, M. Class-based construction of a verb lexicon. AAAI/IAAI, 2000. 691-696.

KLEIN, D. & MANNING, C. D. Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, 2003. Association for Computational Linguistics, 423-430.

KLEIN, D. E. & MURPHY, G. L. 2002. Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language,* 47**,** 548-570.

KUCERA, H. & FRANCIS, W. N. 1967. Computational Analysis of Present-Day {A} merican {E} nglish. Brown University Press, Providence, RI.

LANDAUER, T. K., FOLTZ, P. W. & LAHAM, D. 1998. An introduction to latent semantic analysis. *Discourse processes,* 25**,** 259-284.

LEACOCK, C., MILLER, G. A. & CHODOROW, M. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics,* 24**,** 147-165.

LEE, M. C., CHANG, J. W. & HSIEH, T. C. 2014. A Grammar-Based Semantic Similarity Algorithm for Natural Language Sentences. *The Scientific World Journal,* 2014.

LEE, M. D., PINCOMBE, B. & WELSH, M. B. 2005. An empirical evaluation of models of text document similarity. *Cognitive Science*.

LESK, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the 5th annual international conference on Systems documentation, 1986. ACM, 24-26.

LEVIN, B. 1993. *English verb classes and alternations: A preliminary investigation*, University of Chicago press.

LI, L., ZHOU, Y., YUAN, B., WANG, J. & HU, B.-Y. Sentence similarity measurement based on shallow parsing. Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on, 2009. IEEE, 487-491.

LI, Y., BANDAR, Z. A. & MCLEAN, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *Knowledge and Data Engineering, IEEE Transactions on,* 15**,** 871-882.

LI, Y., MCLEAN, D., BANDAR, Z. A., O'SHEA, J. D. & CROCKETT, K. 2006. Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on,* 18**,** 1138-1150.

LIN, D. Automatic retrieval and clustering of similar words. Proceedings of the 17th international conference on Computational linguistics-Volume 2, 1998a. Association for Computational Linguistics, 768-774.

LIN, D. An information-theoretic definition of similarity. ICML, 1998b. 296-304.

LIU, X.-Y., ZHOU, Y.-M. & ZHENG, R.-S. Measuring semantic similarity in WordNet. Machine Learning and Cybernetics, 2007 International Conference on, 2007. IEEE, 3431-3435.

LIU, X.-Y., ZHOU, Y.-M. & ZHENG, R.-S. Measuring semantic similarity within sentences. 2008 International Conference on Machine Learning and Cybernetics, 2008. 2558-2562.

LUTFI, M. K. 1948. Needed in Egyptian readers to increase their value as media of instruction. PhD thesis, University of Chicago, Chicago.

MAAMOURI, M. & BIES, A. 2010. The Penn Arabic Tree Bank. *Computational Approaches to Arabic Script-Based Languages: Current Implementations in Arabic NLP. CSLI NLP Series*.

MANNING, C. D. & SCHÜTZE, H. 1999. *Foundations of statistical natural language processing*, MIT press.

MARSH, J. E., HUGHES, R. W. & JONES, D. M. 2008. Auditory distraction in semantic memory: A process-based approach. *Journal of Memory and Language,* 58**,** 682-700.

MCCARTHY, J. J. 1981. A prosodic theory of nonconcatenative morphology. *Linguistic inquiry***,** 373-418.

MCCORD, M. C. & CAVALLI-SFORZA, V. An arabic slot grammar parser. Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, 2007. Association for Computational Linguistics, 81-88.

MENG, L., HUANG, R. & GU, J. 2014. Measuring Semantic Similarity of Word Pairs Using Path and Information Content. *International Journal of Future Generation Communication & Networking,* 7.

MERHBEN, L., ZOUAGHI, A. & ZRIGUI, M. 2012. Lexical Disambiguation of Arabic Language: An Experimental Study. *Polibits***,** 49-54.

MERHBENE, L., ZOUAGHI, A. & ZRIGUI, M. 2013. A Semi-Supervised Method for Arabic Word Sense Disambiguation Using a Weighted Directed Graph. International Conference on Natural Language Processing, 1027-1031.

MIHALCEA, R., CORLEY, C. & STRAPPARAVA, C. Corpus-based and knowledge-based measures of text semantic similarity. AAAI, 2006. 775-780.

MILLER, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM,* 38**,** 39-41.

MILLER, G. A. & CHARLES, W. G. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes,* 6**,** 1-28.

MILLER, G. A., LEACOCK, C., TENGI, R. & BUNKER, R. T. A semantic concordance. Proceedings of the workshop on Human Language Technology, 1993. Association for Computational Linguistics, 303-308.

MIRMAN, D. & GRAZIANO, K. M. 2012. Individual differences in the strength of taxonomic versus thematic relations. *Journal of experimental psychology: General,* 141**,** 601.

MITCHELL, J. & LAPATA, M. Vector-based Models of Semantic Composition. ACL, 2008. Citeseer, 236-244.

MOUSSER, J. A Large Coverage Verb Taxonomy for Arabic. LREC, 2010.

NAVIGLI, R. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR),* 41**,** 10.

NG, H. T., WANG, B. & CHAN, Y. S. Exploiting parallel texts for word sense disambiguation: An empirical study. Proceedings of the 41st Annual Meeting on

Association for Computational Linguistics-Volume 1, 2003. Association for Computational Linguistics, 455-462.

NILES, I. & PEASE, A. Linking lexicons and ontologies: mapping WordNet to the Suggested Upper Merged ontology, Proceedings of the IEEE International Conference on IKE, 2003. 412-416.

NIVRE, J., HALL, J., NILSSON, J., CHANEV, A., ERYIGIT, G., KÜBLER, S., MARINOV, S. & MARSI, E. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering,* 13**,** 95-135.

NOAH, S. A., AMRUDDIN, A. Y. & OMAR, N. 2007. Semantic similarity measures for malay sentences. *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers.* Springer.

O'SHEA, J. 2010. A framework for applying short text semantic similarity in goal-oriented conversational agents. *Computing and Mathematics***,** 413.

O'SHEA, J., BANDAR, Z. & CROCKETT, K. 2013. A new benchmark dataset with production methodology for short text semantic similarity algorithms. *ACM Transactions on Speech and Language Processing (TSLP),* 10**,** 19.

O'SHEA, J., BANDAR, Z., CROCKETT, K. & MCLEAN, D. 2010. Benchmarking short text semantic similarity. *International Journal of Intelligent Information and Database Systems,* 4**,** 103-120.

O'SHEA, K., BANDAR, Z. & CROCKETT, K. 2009. Towards a new generation of conversational agents based on sentence similarity. *Advances in Electrical Engineering and Computational Science.* Springer.

O'SHEA, J., BANDAR, Z., CROCKETT, K. & MCLEAN, D. 2008. A comparative study of two short text semantic similarity measures. *Agent and Multi-Agent Systems: Technologies and Applications.* Springer.

O'SHEA, K., BANDAR, Z. & CROCKETT, K. A novel approach for constructing conversational agents using sentence similarity measures. Proceedings of the World Congress on Engineering, 2008.

O'SHEA, K., BANDAR, Z. & CROCKETT, K. 2010. A conversational agent framework using semantic analysis. *International Journal of Intelligent Computing Research (IJICR),* 1.

OLIVA, J., SERRANO, J. I., DEL CASTILLO, M. D. & IGLESIAS, Á. 2011. SyMSS: A syntax-based measure for short-text semantic similarity. *Data & Knowledge Engineering,* 70**,** 390-405.

OPPENHEIM, A. N. 1992. *Questionnaire design, interviewing and attitude measurement,* Bloomsbury Publishing.

OSATHANUNKUL, K. 2014. Semantic similarity framework for Thai conversational agents. *Computing and Mathematics***,** Manchester Metropolitan University.

PARK, E.-K., RA, D.-Y. & JANG, M.-G. 2005. Techniques for improving web retrieval effectiveness. *Information processing & management,* 41**,** 1207-1223.

PARKER, R., DAVID, G., KE CHEN, JUNBO, K. & KAZUAKI, M. 2009. Arabic Gigaword corpus fourth edition. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T30.

PEASE, A., NILES, I. & LI, J. The suggested upper merged ontology: A large ontology for the semantic web and its applications. Working notes of the AAAI-2002 workshop on ontologies and the semantic web, 2002.

PEDERSEN, T., BANERJEE, S. & PATWARDHAN, S. 2005. Maximizing semantic relatedness to perform word sense disambiguation. *University of Minnesota Supercomputing Institute Research Report UMSI*, 25**,** 2005.

PIRRÓ, G. 2009. A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering*, 68**,** 1289-1308.

POON, H. & DOMINGOS, P. Joint inference in information extraction. AAAI, 2007. 913-918.

QUITREGARD, D. 1994. *Arabic Key Words: The Basic 2000-word Vocabulary Arranged by Frequency in a Hundred Units: with Comprehensive English and Transliterated Arabic Indexes*, Oleandes Press.

RADA, R., MILI, H., BICKNELL, E. & BLETTNER, M. 1989. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on*, 19**,** 17-30.

RESNIK, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 448–453.

RESNIK, P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research, 11, 95–130.

RESNIK, P. & DIAB, M. 2000. Measuring verb similarity. DTIC Document.

RODRÍGUEZ, H., FARWELL, D., FERRERES, J., BERTRAN, M., ALKHALIFA, M. & MARTÍ, M. A. Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. LREC, 2008.

ROGET, P. M. 2008. *Roget'S International Thesaurus, 3/E***, Oxford and IBH Publishing.

RUBENSTEIN, H. & GOODENOUGH, J. B. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8**,** 627-633.

RYDING, K. C. 2005. *A reference grammar of modern standard Arabic*, Cambridge University Press.

SALEM, Y. 2009. *A generic framework for Arabic to English machine translation of simplex sentences using the Role and Reference Grammar linguistic model.* Institute of Technology Blanchardstown Dublin, Ireland.

SAWALHA, M. S. S. 2011. *Open-source resources and standards for Arabic word structure analysis: Fine grained morphological analysis of Arabic text corpora*, University of Leeds.

SCHULER, K. K. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon.

SIBAWAIHI, ABU BISHR AMR. 1966. Al-Kitab. Cairo, Egypt: Dar al-Qalam.

SINCLAIR, J. 2001. *Collins COBUILD English dictionary for advanced learners*, HarperCollins.

SINHA, R. S. & MIHALCEA, R. Unsupervised Graph-basedWord Sense Disambiguation Using Measures of Word Semantic Similarity. ICSC, 2007. 363-369.

SLEATOR, D. D. & TEMPERLEY, D. 1995. Parsing English with a link grammar. *arXiv preprint cmp-lg/9508004*.

SMART, J. 1992. Teach Yourself Arabic: the complete course for beginners, NTC Publishing Group, USA.

SOLER, S. & MONTOYO, A. 2002. A proposal for WSD using semantic similarity. *Computational Linguistics and Intelligent Text Processing*. Springer.

SULEIMAN, M. 1990. Sibawaihi's' Parts Of Speech'According To Zajjaji: A New Interpretation. *Journal of Semitic Studies***,** 245-263.

SULEIMAN, S. M. 1989. On the pragmatics of subject-object preposing in standard Arabic. *Language Sciences,* 11**,** 215-235.

TLILI-GUIASSA, Y. 2006. Hybrid method for tagging Arabic text. *Journal of Computer science,* 2**,** 245.

TOUTANOVA, K., KLEIN, D., MANNING, C. D. & SINGER, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, 2003. Association for Computational Linguistics, 173-180.

TOUTANOVA, K. & MANNING, C. D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13, 2000. Association for Computational Linguistics, 63-70.

VALCOURT, G. & WELLS, L. 1999. *Mastery: A university word list reader,* University of Michigan Press.

VAN OVERSCHELDE, J. P., RAWSON, K. A. & DUNLOSKY, J. 2004. Category norms: An updated and expanded version of the norms. *Journal of Memory and Language,* 50**,** 289-335.

WEHR, H. 1979. *A dictionary of modern written Arabic:(Arab.-Engl.),* Otto Harrassowitz Verlag.

WIEMER-HASTINGS, P. & WIEMER, P. Adding syntactic information to LSA. Proceedings of the 22nd Annual Meeting of the Cognitive Science Society, 2000.

WIGHTWICK, J., GAAFAR, M. & GAAFAR, M. 2007. Mastering Arabic 1, Palgrave Macmillan.

WIGHTWICK, J., GAAFAR, M. & GAAFAR, M. 2009. Mastering Arabic 2, Palgrave Macmillan.

WILKS, Y., FASS, D., GUO, C.-M., MCDONALD, J. E., PLATE, T. & SLATOR, B. M. 1990. Providing machine tractable dictionary tools. *Machine translation,* 5**,** 99-154.

WRIGHT, W. 1896/2005. A Grammar of the Arabic Language. Cambridge: Cambridge University Press.

WU, Z. & PALMER, M. Verbs semantics and lexical selection. Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994. Association for Computational Linguistics, 133-138.

YAROWSKY, D. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd annual meeting on Association for Computational Linguistics, 1995. Association for Computational Linguistics, 189-196.

YOON, C., FEINBERG, F., HU, P., GUTCHESS, A. H., HEDDEN, T., CHEN, H.-Y. M., JING, Q., CUI, Y. & PARK, D. C. 2004. Category norms as a function of culture and age: comparisons of item responses to 105 categories by american and chinese adults. *Psychology and aging,* 19**,** 379.

Appendices

Appendix 1

This appendix contains examples (in Arabic and English) of experimental materials used in the first experiment of the Arabic noun dataset creation methodology which is the experiment of the construction of the set of Arabic noun pairs.

Appendix 1.1 Ethics statement

نرغب في الحصول على موافقتك للمشاركة في دراسة علمية عن التشابة الدلالي للغة العربية.

بعد الموافقة على المشاركة في هذه الدراسة سوف يطلب منك كتابة مجموعة من أزواج الكلمات العربية بينهم تشابه دلالي في المعنى. سوف تزود باستمارة تحتوي على مجموعة من الكلمات تستخدمها في كتابة ازواج الكلمات العربية. مجموعة الكلمات هذه هي كلمات شائعة (تستخدم بكثرة في الحياة اليومية) ولا تحتوي على اي كلمة تسبب لك مشكلة.

في نهاية الاستبانة سوف يطلب منك كتابة بعض المعلومات الشخصية والتي تتضمن الاسم، العمر، المؤهل، من اي قطر عربي انت وكذلك تأكيد على ان اللغة العربية هي لغتك الام. نحن نحتاج هذه المعلومات فقط للتأكيد ان المجموعة المشاركة في هذه التجربة مثلا من مختلف الاقطار العربية ومن فئات عمرية مختلفة وهكذا. هذه المعلومات سوف تبقى معنا لمدة لا تتجاوز 3 شهور بعد نشر النتائج الاولى.

مجموعة ازواج الكلمات التي تكتبها سوف يتم فصلها عن معلوماتك الشخصية وذلك لغرض الاحتفاظ بها بشكل دائم للاستفادة منها في دراسات اخرى.

معلوماتك الشخصية سوف لن تكشف او تعطى لاي جهة ليست لها صلة بالبحث ولكن نتائج البحث سوف يتم نشرها دوليا.

Appendix 1.2 instruction sheet

نشكر لك تطوعك للمشاركة في هذه الدراسة .

الرجاء قراءة المعلومات أدناه قبل البدء بالاستبانة .

في هذه التجربة ، نرغب في الحصول على مساعدتك لكتابة قائمة من أزواج الكلمات العربية ولهذا الغرض سوف تزود بأربع إستمارات تتضمن :

- **<u>استمارة الكلمات العربية</u>** التي تحتوي على 56 كلمة عربية مختلفة ممثلة في مجموعتين(**أ** و **ب**). كل مجموعة تتكون من 28 كلمة .

المطلوب منك استخدام هذه الاستمارة لكتابة ازواج الكلمات عن طريق **اختيار كلمة واحدة من المجموعة – أ – وكلمة أخرى من المجموعة – ب –** .

- **<u>استمارتا تسجيل</u>** تستخدم لكتابة أزواج الكلمات العربية طبقاً لنوع التشابه في المعنى بين الكلمتين المختارتين.

ما المقصود بالتشابه في المعنى؟

يجب ان تنظر الى زوج الكلمة (نقصد الكلمتين اللتين يتم اختيارهما من استمارة الكلمات العربية) وتسأل نفسك

- مدى تقارب هاتين الكلمتين لتعطي المعنى نفسه ؟ أو
- مدى تقارب هاتين الكلمتين لجعلك تشعر او تعتقد أنهما يعطيان المعنى نفسه ؟

والمطلوب منك كتابة قائمتين بأزواج الكلمات طبقاً لنوع التشابه في المعنى :

- قائمة التشابه العالي في المعنى التي تحتوي على أزواج الكلمات التي تعطي تشابهاً ما بين **علاقة قوية و متطابقة** (مترادفة) في المعنى.
- قائمة التشابه المتوسط في المعنى التي تحتوي على أزواج الكلمات التي تعطي تشابهاً متوسطاً قليلاً (**اي ارتباط ضمني في المعنى**) وتشابهاً متوسطاً عالياً (**اي متشابه كثيرا في المعنى**).

- **<u>استمارة البيانات الشخصية</u>** للحصول على بعض البيانات الاولية عنك التي تتضمن الاسم والعمر والجنس والتحصيل العلمي بالاضافة الى التأكيد على إن اللغة العربية هي لغتك الام وتتحدث بها منذ الولادة .

من فضلك لا تكرر كتابة ازواج الكلمات في نفس الاستمارة او بين استمارتي التسجيل.

نـــود منـــك أن تفكـــر مليـــا فـــي كـــل زوج مـــن الكلمـــات التـــي ســـوف تكتبهـــا فـــي أســـتمارتي التســـجيل. كمـــا يمكنـــك تكـــرار اســـتخدام بعـــض الكلمـــات فـــي اســـتمارة الكلمـــات العربيـــة لتكـــوين ازواج جديـــدة اذا كنـــت تعتقـــد انهـــا مطابقـــة لنـــوع التشـــابه فـــي المعنـــى فـــي اســـتمارة التســـجيل التي تملؤها.

Appendix 1.2 Instruction Sheet (English copy)

Please read before you start performing the task.

Thank you for volunteering to participate in this study.

In this experiment we would like you to help us for constructing a list of Arabic word pairs. You will be supplied with 4 sheets which include:

- **The sheet of Arabic nouns** contains 56 different Arabic words (nouns) represented in two columns (A and B). Each column has 28 Arabic words.

  You are requested to use the Arabic nouns sheet for writing a list of Arabic word pairs by selecting **one** word from group A and **one** word from group B.

- **Two recording sheets** to write two lists of word pairs according to amount of similarity of meaning.

  What do we mean by similarity of meaning?

  You should look at the word pair (two words you will select them from the theme words sheet) and ask yourself

  - How close do these two words come to meaning the same thing?
  - How close do they come to making you feel or believe the same thing?

  Or

  - How close do they come to making you do the same thing?

  You are requested to write two lists of word pairs according to amount of similarity of meaning:

  - The high similarity of meaning list contains word pairs between **strongly related** in meaning and **identical** in meaning.
  - The Medium similarity of meaning list contains word pairs between **vaguely similar** in meaning and **very much alike** in meaning.

- **Personal information sheet** to complete a few details about you. These are your name, age, gender etc.

We would like you to think carefully about each word pair you will write it in a recording sheet. Please note that you can select any word from the column A more than once with different words from the column B to create new word pairs.

Please do not write the same word pair more than once in the same sheet or between different sheets.

Appendix 1.3 Medium Similarity Word Pairs Recording Sheet (Arabic noun dataset, Experiment 1: Constructing the set of Arabic noun pairs)

- مـــن فضـــلك اســـتخدم اســـتمارة الكلمـــات العربيـــة لتكـــوين أزواج الكلمـــات ذات التشـــابه المتوسط في المعنى. يرجى كتابة الازواج في الحقول المخصصة لها أدناه.
- يرجـــى ملاحظـــة ان أزواج الكلمـــات دائمـــاً تحتـــوي علـــى كلمـــة واحـــدة مـــن المجموعـــة – **أ-** وكلمة أخرى من المجموعة – **ب -** .
- المقصـــود بالتشـــابه المتوســـط فـــي المعنـــى ان تعطـــي أزواج الكلمـــات تشـــابهاً متوســـطاً قلـــيلاً **(اي ارتبـــاط ضـــمني فـــي المعنـــى)** وتشـــابهاً متوســـطاً عالـــياً **(اي متشـــابه كثيـــرا فـــي المعنى)**.
- الرجـــاء كتابـــة 32 زوجـــاً مـــن أزواج الكلمـــات ذات التشـــابه المتوســـط فـــي المعنـــى حيـــث سيتم تجـاهل جميـع الاستبيـانات الغيـر مكتملـة.

.1	.1
.2	.2
.3	.3
.4	.4
.5	.5
.6	.6
.7	.7
.8	.8
.9	.9
.28	.28
.29	.29
.30	.30
.31	.31
.32	.32

Appendix 1.3 Medium Similarity Word Pairs Recording Sheet (English copy)

- Please, use the Arabic nouns sheet to create word pairs that have a Medium similarity of meaning.
- Please note that the word pairs always contain one word from group A and one word from group B.
- The medium similarity of meaning means that the word pairs have a similarity between vaguely similar in meaning and very much alike in meaning.
- Please, write 32 word pairs for a medium similarity of meaning since all uncompleted questionnaires must be ignored.

1.	1.
2.	2.
3.	3.
4.	4.
5.	5.
6.	6.
7.	7.
8.	8.
9.	9.
28.	28.
29.	29.
30.	30.
31.	31.
32.	32.

Appendix 1.4 Personal Information Sheet (for all experiments of the noun and verb datasets)

الاسم:

العمر:

الجنس: **أنثى** **ذكر**

الدولــة :

التحصيل العلمي:
(متضمنا موضوع الدراسة)

هل اللغة العربية هي لغتك الام : **نعم** **كلا**

التوقيـع :

Appendix 1.4 Personal Information Sheet

Name:

Age:

Gender: Female **Male**

Your country:

Your highest education qualification (including subject):

Confirmation that you are a native Arabic speaker*: Yes **No**

Signature:

Appendix 2

This appendix contains examples of experimental materials used in the second experiment of the Arabic noun dataset creation methodology which is the experiment of the collection of the human similarity ratings.

Appendix 2.1 Instruction Sheet

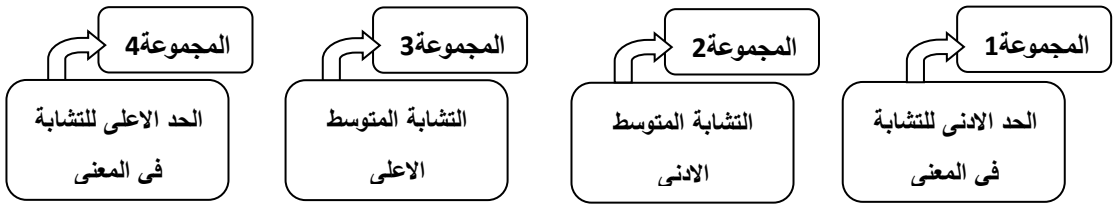
نشكر لك تطوعك للمشاركة في هذه الدراسة .

الرجاء قراءة المعلومات أدناه قبل البدء بالاستبانة .
تستطيع الانسحاب قبل البدء بالاجابة او في اي مرحلة من مراحل الاستبانة.

- في هذه الدراسة ستزود بمغلف يحتوي على مجموعة من البطاقات مرتبة ترتيبا عشوائيا وكل بطاقة كتب عليها كلمتين. نرغب في الحصول على مساعدتك لاعطاء مقدار التشابة في المعنى بين هاتين الكلمتين. لهذا ستزود باستمارة تسجيل لكتابة مقدار التشابة لكل زوج من الكلمات المكتوبة في البطاقات .

  المقصود بالتشابه في المعنى مدى تقارب هاتين الكلمتين لتعطي المعنى نفسه ؟ أو
  مدى تقارب هاتين الكلمتين لجعلك تشعر او تعتقد أنهما يعطيان المعنى نفسه ؟

- الان من فضلك افتح المغلف ورتب البطاقات التي تجدها داخله في اربعة مجاميع طبقا لنوع التشابة في المعنى وكالاتي:
  - مجموعة الحد الاعلى للتشابة والتي تحتوي على ازواج الكلمات المتطابقة (المترادفة) في المعنى و الازواج التي بينها علاقة قوية في المعنى.
  - مجموعة الحد الادنى للتشابة والتي تحتوي على ازواج الكلمات التي لا يوجد بينها ارتباط في المعنى على الاطلاق.
  - المجموعتان الاخريان تحتويان على ازواج الكلمات التي تقع بين الحد الاعلى والحد الادنى للتشابة (الكلمات التي بينها تشابة ضمني في المعنى توضع في مجموعة التشابة المتوسط الادنى بينما الكلمات التي بينها تشابة واضح اكثر من كونه ضمني يوضع في مجموعة التشابة المتوسط الاعلى).



- بالنسبة لعدد البطاقات في كل مجموعة الامر متروك لقرارك بشأن كل بطاقة.

- الان من فضلك اقرأ البطاقات في كل مجموعة بعناية. اذا غيرت رأيك بالنسبة الى المجموعة التي يجب ان تكون بها اي بطاقة الرجاء نقل البطاقة الى المجموعة الاخرى.

- من فضلك اكتب مقدار التشابة لكل زوج عن طريق كتابة رقم بين (0.0 – 4.0) في استمارة التسجيل كما موضح ادناه:
  - 0.0 بالنسبة للكلمات التي لا يوجد ارتباط بينها في المعنى.
  - 1.0 الكلمات التي بينها تشابة ضمني في المعنى.
  - 2.0 الكلمات التي بينها تشابة واضح (اكثر من كونه ضمني) في المعنى.
  - 3.0 الكلمات التي بينها علاقة قوية في المعنى.
  - 4.0 الكلمات المترادفة او المتطابقة في المعنى.

- من فضلك لا تكتب مقدار اعلى من 4.0 او اقل من 0.0. تستطيع استخدام رقم عشري واحد بعد الفارزة لكتابة درجة تشابة دقيقة (مثلا اذا كنت تعتقد ان التشابة بين الكلمتين في البطاقة ما بين 2.0 و 3.0 يمكنك كتابة الرقم 2.5).

Appendix 2.1 Instruction Sheet

Thank you for volunteering to participate in this study.

You can withdraw before beginning the questionnaire or at any point while performing the questionnaire.

You are supplied with an envelope containing 70 cards (each card contains different Arabic word pair) and a recording sheet to write your ratings on. In this experiment, we would like you to help us by reading each card carefully and thinking about the similarity of meaning of the two words written on it.

What do we mean by similarity of meaning?

You should look at the word pair on each card and ask yourself

How close do these two words come to meaning the same thing or making you believe the same thing?

- Now please sort the 70 cards into four groups according to amount of similarity of meaning.
  - The high similarity of meaning group contains word pairs between strongly related in meaning and identical in meaning.
  - Minimum similarity of meaning group contains word pairs unrelated in meaning.
  - Two medium similarity of meaning groups contain word pairs vaguely similar in meaning for low medium similarity of meaning and very much alike in meaning for high medium similarity of meaning.

- The number of cards in each group is based on your judgement on each card.
- Please check the cards in each group carefully; you may change a word pair from group to other in this stage.
- Please rate each word pair according to amount of similarity of meaning by writing one of the 5 points rating scales as follows.

- 0.0 The word pairs are unrelated in meaning.
- 1.0 The word pairs are vaguely similar in meaning.
- 2.0 The word pairs are very much alike in meaning.
- 3.0 The word pairs are strongly related in meaning.
- 4.0 The word pairs are identical in meaning.

- Please do not write values greater than 4.0 or less than 0.0. You can also use the first decimal place to assign an accurate degree of similarity (for example, if you think the similarity of word pair between 2 and 3 you can assign value like 2.5)". Also, you may rate more than one word pair with the same value.

Appendix 2.2 Similarity Rating Recording Sheet (Arabic noun dataset, Experiment 2: collection of the human similarity ratings)

- من فضلك اكتب مقدار التشابة في المعنى لكل زوج من ازواج الكلمات.
- يرجى ملاحظة ان مقدار التشابة يكون بين 0.0 (الحد الادنى للتشابة) و 4.0 (الحد الاعلى للتشابة)
- من فضلك لا تكتب مقدار اعلى من 4.0 او اقل من 0.0. كما يمكنك كتابة نفس مقدار التشابة لاكثر من زوج.
- يمكنك استخدام رقم عشري واحد بعد الفارزة لكتابة مقدار تشابة دقيق.

ز 01	
ز 02	
ز 03	
ز 04	
ز 05	
ز 06	
ز 07	
ز 10	
ز 11	
ز 12	
ز 13	
ز 14	
ز 15	
ز 16	
ز 20	
ز 21	
ز 22	
ز 23	
ز 24	
ز 25	
ز 26	
ز 30	

ز 66	
ز 70	
ز 71	
ز 72	
ز 73	
ز 74	
ز 75	
ز 76	
ز 80	
ز 81	
ز 82	
ز 83	
ز 84	
ز 85	
ز 86	
ز 90	
ز 91	
ز 92	
ز 93	
ز 94	
ز 95	
ز 96	

Appendix 2.2 Similarity Rating Recording Sheet

- Please, enter a rating for the similarity of meaning of each word pair.
- Please note that the rating scale runs from 0.0 (minimum similarity) to 4.0 (maximum similarity).
- Please do not write values greater than 4.0 or less than 0.0. Also, you may assign the same value to more than one pair.
- You can use the first decimal place to write an accurate degree of similarity.

WP 01	
WP 02	
WP 03	
WP 04	
WP 05	
WP 06	
WP 07	
WP 10	
WP 11	
WP 12	
WP 13	
WP 14	
WP 15	
WP 16	
WP 20	
WP 21	
WP 22	
WP 23	
WP 24	
WP 25	
WP 26	
WP 30	

WP 66	
WP 70	
WP 71	
WP 72	
WP 73	
WP 74	
WP 75	
WP 76	
WP 80	
WP 81	
WP82	
WP 83	
WP 84	
WP 85	
WP 86	
WP 90	
WP 91	
WP 92	
WP 93	
WP 94	
WP 95	
WP 96	

Appendix 2.3 A Sample of Arabic word pair card (Arabic noun dataset, Experiment 2: collection of the human similarity ratings)

A sample of Arabic word pair card.

WP 50	
Word 1	Glass
Word 2	Tumbler

زك 50	
الكلمة 1	**كـــأس**
الكلمة 2	**قـــدح**

Appendix 3

This appendix contains examples of experimental materials used in the first experiment of the Arabic verb dataset creation methodology which is the experiment of the construction of the set of Arabic medium similarity verb pairs. Generation this set required conducting two experiments (1 & 2) to create two lists of synonyms. The following materials used in the experiment 1 and 2.

Appendix 3.1 Instruction Sheet

نشكر لك تطوعك للمشاركة في هذه الدراسة .

الرجاء قراءة المعلومات أدناه قبل البدء بالاستبانة .

في هذه التجربة ، نرغب في الحصول على مساعدتك لكتابة فعلين لكل فعل في قائمة الافعال العربية التي ستزود بها.

هذه التجربة ليست اختبار لك باي شكل من الاشكال ولكن للحصول على بيانات يمكن ان تستخدم في تجارب مستقبلية لاحتساب التشابه بين ازواج من الافعال العربية.

سوف تجهز باستمارتين تتضمن :

- **استمارة الافعال العربية** تحتوي على قائمة من 23 فعل عربي. المطلوب منك كتابة فعلين لكل فعل في القائمة. الافعال التي ستكتبها يجب ان تعطي نفس المعنى للفعل الاصلي او قريبة جدا في المعنى من الفعل الاصلي.
- **استمارة البيانات الشخصية** للحصول على بعض البيانات الاولية عنك التي تتضمن الاسم والعمر والجنس والتحصيل العلمي بالاضافة الى التأكيد على إن اللغة العربية هي لغتك الام وتتحدث بها منذ الولادة .

نود منك أن تفكر مليا في كل فعل من الافعال التي سوف تكتبها. من فضلك لا تكتب اكثر من فعلين لكل فعل من الافعال 23 في استمارة الافعال العربية.

الرجاء كتابة الافعال بخط واضح كي يتسنى لنا قراءته وادخاله الى الحاسوب.

من فضلك لا تكرر كتابة الفعل الاصلي ولا تكتب نفس الفعل مرتين كجواب.

من فضلك يجب ان تكتب فعلين لكل فعل من الافعال 23 في استمارة الافعال العربية وذلك لان الاستبيان غير المكتمل يجب ان يتم تجاهله.

Appendix 3.1 Instruction Sheet

Please read before you start doing the task.

Thank you for volunteering to participate in this study.

We would like you to assist us in this experiment by writing two verbs for each verb in the list of Arabic verbs that we will supply.

This experiment does not test you in any way; it is to generate data that can be employed in future experiments for measuring the similarity between two verbs.
You will be supplied with 2 sheets which include:

- **The sheet of Arabic verbs** contains a list of 23 different Arabic verbs. You are requested to write two verbs for each verb in this list. The verbs that you will write mean the same as or very close in meaning to the original Arabic verb.
- **Personal information sheet** to complete a few details about yourself. These are your name, age, gender etc.

We would like you to think carefully about each verb you will write. Please do not write more than two verbs for each verb in the list of Arabic verbs.

Please print the verbs that you will write in clear handwriting so it is easier for us to read and type them into the computer.

Please do not write the original verb and do not write the same verb twice as an answer.

Please write two verbs for each of the 23 verbs in the list of Arabic verbs because all uncompleted questionnaires must be ignored.

Appendix 3.2 Verb Recording Sheet (Arabic verb dataset, construction of the set of Arabic medium similarity verb pairs, Experiments 1: creation of the lists of synonyms)

- الرجاء كتابة فعلين لكل فعل من الافعال الاصلية في القائمة ادناه والتي يمكن ان تحل محل الفعل الاصلي في جملة. هذا يعني ان الافعال التي ستكتبها يجب ان تعطي نفس المعنى للفعل الاصلي او قريبة جدا في المعنى من الفعل الاصلي.
- الرجاء كتابة فعلين لكل فعل من الافعال 23 في القائمة ادناه.

	الفعل الاصلي	الفعل 1	الفعل 2
1	تمكن		
2	تضمن		
3	أعتبر		
4	ظهر		
5	وجب		
6	حصل		
7	اتصل		
8	أزدحم		
9	إرتفع		
10	تمنى		
11	حدث		
12	وجد		
13	أغنى		
14	غادر		
15	تسرب		
16	استمر		
17	حاول		
18	عين		
19	صرح		
20	وافق		
21	أعطى		
22	وصل		
23	ملأ		

Appendix 3.2 Verb Recording Sheet

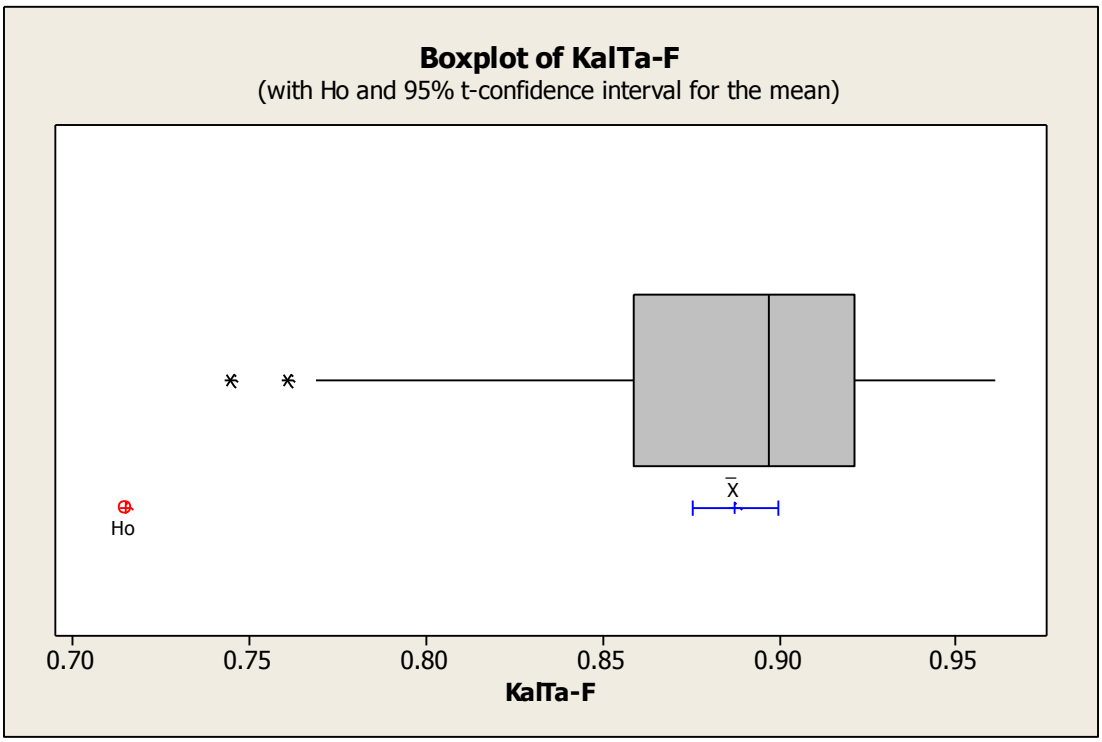
- For each original verb, please write two verbs that could be used in its place in a sentence, i.e. means the same or very close in meaning.

- Please, write two verbs for each of the 23 Arabic verbs listed below.

	Original verbs	Verb1	Verb2
1	Be capable		
2	Include		
3	Consider		
4	Appear		
5	Be obligatory		
6	Obtain		
7	Contact		
8	Crowd		
9	Rise		
10	Hope		
11	Happen		
12	Find		
13	Enrich		
14	Depart		
15	Leak		
16	Continue		
17	Try		
18	Appoint		
19	Declare		
20	Approve		
21	Give		
22	Arrive		
23	Fill		

Appendix 4

This appendix is summarized the result of the one sample t-test with confidence interval plot which used to compare between a single correlation obtained by KalTa-F without Root (Arabic verb measure) and the average of the correlation coefficients on the evaluation dataset.



Appendix 5

This appendix contains a list of new Arabic categories (20 categories) created in Arabic short text dataset which used in the stage of the selection of the set of Arabic words.

	Categories Names	اسماء الفئات العربية
1	Military title	لقب عسكـري
2	Part of human body	اجزاء من جسم الانسان
3	An occupation	مهنـــة
4	Four footed animals	حيوانات تمشي على اربع
5	Insect	حشرات
6	Fish	سمك
7	Diseases	امراض
8	snake	افعى
9	Fruit	فواكهة
10	Tree	اشجار
11	Vegetable	خضروات
12	Flower	زهور
13	Metal	معادن
14	Type of reading material	مادة للقراءة
15	Building for religious services	مبنى للخدمات الدينية
16	Weapon	سلاح
17	Weather phenomenon	ظاهرة مناخية
18	Non-alcoholic beverage	مشروبات غير كحولية
19	Crime	جريمة
20	Part of building	اجزاء من المباني

Appendix 6

This appendix lists the Arabic themes that investigated for the experiment of creation of the Arabic short text database.

Teach yourself Arabic (smart 1992)

1. The Muslim festivals
2. Islamic calendar
3. At the airport
4. Arabic social structure
5. Greeting, polite phrases and forms of address
6. Given orders
7. Islamic conquests

Mastering Arabic part 1 (Wightwick and Gaafar 2007)

1. The family
2. Jobs
3. Countries and people
4. Describing places
5. Where is it?
6. What happened yesterday?
7. Eating and drinking
8. Comparing things
9. Future plans
10. Months of the year
11. Wish you where here

Mastering Arabic part 2 (Wightwick and Gaafar 2009)

1. Sport and leisure
2. Travel and tourism
3. Food and drinking
4. Cloths and colours
5. Work and routine

6. Education and training
7. News and media
8. House and home
9. Climate and environment
10. Health and happiness
11. Arts and cinema

Appendix 7

This appendix contains examples of experimental materials used in the experiment of the Arabic short text dataset creation methodology which is the experiment of the creation of the database of 1088 Arabic short texts.

Appendix 7.1 Instructions Sheet

نشكر لك تطوعك للمشاركة في هذه الدراسة.

الرجاء قراءة المعلومات أدناه قبل البدء بالاستبانة.
تستطيع الانسحاب قبل البدء بالاجابة او في اي مرحلة من مراحل الاستبانة.

معظم صفحات الاستبانة تتضمن تعليمات لكتابة جمل لمجموعة من الكلمات المحددة.

ادناه مثال على نوع التعليمات التي سوف تجدها في صفحات الاستبانة:

مثال: الرجاء كتابة **جملتين** بخط واضح كل جملة **ما بين 10 الى 20** كلمة طولا تحوي كل منها على الاسم **مستشفى**

الجملتين التي تكتبها يجب ان تكون ضمن واحدة من الصيغ (الاشكال) التالية

- استفهام (سؤال)
- تعليمات
- تعابير
- خبر او بيان
- التزام
- تصريح

توجد استمارة منفصلة تشرح او توضح كل صيغة من الصيغ اعلاه (صفحة 2)

ملاحظات مهمة:

- اذا كانت **الكلمة** المراد كتابة جملتين لها تحمل اكثر من معنى يمكنك اختيار معنى واحد ولكن من فضلك استخدم نفس المعنى في كتابة الجملتين. كما يجب الالتزام بنوع الكلمة المعطى في التعليمات من حيث كونها فعل, اسم, صفة او ظرف.

- هـذه الدراسـة ليسـت لاختبـارك بـاي شـكل مـن الاشـكال وانمـا تبحـث فـي الاسـتخدام اليـومي للغة العربية الفصحى.

- نرغـب فـي الحصـول علـى مسـاعدتك مـن خـلال كتابـة جمـل طبيعيـة ذات معنـى والتـي ربمـا تسـتخدمها انـت فـي التحـدث او الكتابـة او ربمـا تُسـتخدم مـن قبـل النـاس فـي التواصل معك.

- مـن فضـلك الجمـل التـي تُكتـب يجـب ان تكـون بصـيغة تعليمـات، تعـابير، التـزام، خبـر او تصريح. الرجاء عدم التمسك بصيغة واحدة فقط لكتابة كل جمـل الاستبانة.

- فـي نهايـة الاسـتبانة سـوف يطلـب منـك كتابـة بعـض المعلومـات الشخصـية والتـي تتضـمن الاسـم، العمـر، المؤهـل، مـن اي قطـر عربـي انـت وكـذلك تأكيـد علـى ان اللغـة العربيـة هـي لغتـك الام. نحـن نحتـاج هـذه المعلومـات فقـط للتأكيـد ان المجموعـة المشـاركة فـي هـذه التجربـة مـثلا مـن مختلـف الاقطـار العربيـة وان لـديهم المـام باللغـة العربيـة وهكـذا. هذه المعلومات سوف تبقى معنا لمدة لا تتجاوز 3 شهور بعد نشر النتائج الاولى.

Appendix 7.1 Instructions Sheet

Thank you for volunteering to participate in this study.
Please read before you start performing the task.

You can withdraw before beginning the questionnaire or at any point while performing the questionnaire.

Most of the questionnaire's pages include an instruction to write two short texts using a particular word.

Below an example of this type of instruction:
Please write two short texts in clear handwriting between 10 to 20 words in length, each containing the noun Hospital

Your short texts should be in one of the following forms:

- A question
- An expression
- A declaration
- A commitment
- An instruction
- A statement

There is a separate sheet explains each form (page 2).

If the word that is used to write the short texts have more than one meaning you can select which one to use but please use the same meaning in writing the two short texts and stick to the type of word given in the instructions, if it is Noun, Verb, Adjective or Adverb.

This study is not to measure you for creativity; it is only looking at the everyday use of the Arabic Language.

We would like you to help us by writing natural and meaningful short texts that you might actually say or write, or that other people might use to communicate with you.

Please don't stick to a particular form of short text; they can be questions, statements, instructions, expressions, commitments or declarations.

Appendix 7.2 contains samples extracted from the questionnaire of the experiment of the creation of the database of 1088 Arabic short texts.

الصفحـات الثلاث القادمـة ستكون حول **الصـفة**

الصـفة : كلمـة تـدل على موصوف. مثـال على ذلك **جـبل شامـخ**.
يـمكنك استخـدام اي صـيغة صـحيحة للصفة في الجملة التي تكتبها مثلا
شامخ، شامخة، شامخات، .....

هنـاك بعـض الكلمـات ممكـن ان تسـتخدم كصـفة أو أسـم مثـل (**زائـر**). مـن فضـلك هـذا النـوع مـن الكلمـات يجـب اسـتخدامها كصـفة فـي هـذا الجـزء مـن الاستبانة.
هنـاك بعـض الكلمـات ممكـن ان تسـتخدم كصـفة أو فعـل مثـل (**أوضـح**). مـن فضـلك هـذا النـوع مـن الكلمـات يجـب اسـتخدامها كصـفة فـي هـذا الجـزء مـن الاستبانة.

مـن فضـلك اسـتخدم الصـفة بالصـيغة المعطـاة فـي التعليمـات. مـثلا اذا اعطيـت الصـفة (جميـل) يمكـن اسـتخدامها (جميلـة، جمـيلات) لكـن **لا تسـتخدمها** بصيغة (أجمل او الاجمل) .

The next three pages are about the adjectives

Adjective is a word that is used to describe a noun, for example, lofty mountain.

You can use any valid form of the adjective in the short text you write, for example: ***ShaAmix*** (lofty, for male), ***ShaAmixah*** (lofty, for female), ***ShaAmixAt*** (plural) …..

Some Arabic words such as "visitor" can be used as a noun or adjective. Please use this word as adjective in this part of the questionnaire.

Some Arabic words such as "clearer" can be used as verb or adjective. Please use this this word as adjective in this part of the questionnaire.

Please use the adjective in the form as given in the instruction. For example, if you are given the adjective "beautiful" do not use it as a comparative "more beautiful" or a superlative "most beautiful".

الرجـاء كتابـة **جملتـين** بخـط واضـح كـل جملـة **مـا بـين 10 الـى 20** كلمـة طـولا تحوي كل منها على الصفة **أزرق**

الجملتـين التـي تكتبهـا يجـب ان تكـون ضـمن واحـدة مـن الصـيغ (الاشـكال) التالية

- استفهام (سـؤال)
- تعليمات
- تعابير
- خبر او بيان
- التزام
- تصريح

توجـد استمـارة منفصـلة تشـرح او توضـح كـل صـيغة مـن الصـيغ اعـلاه (صفحة 2)

الجملـة الاولى

_____
_
_____

الجملـة الثانية

_____
_____

Please write two short texts in clear handwriting between **10** to **20** words in length, each containing the Adjective **Blue.**

Your short texts should be in one of the following forms:

- A question
- An expression
- A declaration
- A commitment
- An instruction
- A statement

There is a separate sheet explains each form (page 2).

Short text 1

_____

_____

Short text 2

_____

_____

الرجـاء كتابـة **جملتـين** بخـط واضـح كـل جملـة **مـا بـين 10 الـى 20** كلمـة طـولا تحـوي كـل منهـا علـى الاسـم **سرطـان** ويجـب ان يكـون موضـوع الجملتـين **(الصحـة و السعادة)**

الجملتـين التـي تكتبهـا يجـب ان تكـون ضـمن واحـدة مـن الصـيغ (الاشـكال) التالية

- استفهام (سـؤال)
- تعليمات
- تعابير
- خبر او بيان
- التزام
- تصريح

توجـد استمـارة منفصـلة تشـرح او توضـح كـل صـيغة مـن الصـيغ اعـلاه (صفحة 2)

الجملـة الاولى

_____
_____

الجملـة الثانية

_____
_____

Please write two short texts in clear handwriting between **10** to **20** words in length, on the general topic of **Health and happiness** and each containing the Noun **Cancer.**

Your short texts should be in one of the following forms:

- A question
- An expression
- A declaration
- A commitment
- An instruction
- A statement

There is a separate sheet explains each form (page 2).

Short text 1

_____

_____

Short text 2

_____

_____

Appendix 7.3 Dialogue Act clarification sheet (Creation of the Arabic short text database experiment)

الرجـــاء قـــراءة المعلـــومات ادنـــاه بعنايـــة والتـــي توضـــح او تشـــرح الصـــيغ التـــي تستخدم لكتابة جمل الاستبانة.

- **البيان أو الخبر**: توضيح، وصف، تصنيف، خبر

مثال:

قيـادة الطـائرة أصعـب بكثـير من قيـادة السـيارة.

- **تعليمات:** طلب، أمر، ،ايعاز او تعليمة

مثال: **لا تستخدم الهاتف النقال اثناء قيادة السيارة.**

- **الالتزامات:** الوعود، الضمانات، العقود، التعهدات

مثال: **سـأقود السـيارة حسـب قوانـين السـير البريطـانيـة.**

- **التعابير**: الاعتذار، الشكر، التهاني، الترحيب، التعازي

مثـال: **اقـــدم اعتـــذاري عـــن قيـــادة الحـــزب وذلـــك بســـبب مشـــاكلي الصحـية.**

- **تصريحات** : اعلانات ، تصريحات

مثـال: **قـــررت شـــركة مصـــر للطيـــران تنظـــيم دورات تعليميـــة حـــول مشـــاكل قيادة الطائرات لكوادرها.**

Appendix 7.3 Dialogue Act clarification sheet

Statement:

Statements, explanations, descriptions, classifications.

e.g.

Piloting an airplane is much more difficult than driving a car.

Instruction:

Instructions, requests, orders, commands.

e.g.

Do not use a mobile phone while driving a car.

Commitments:

Promises, guarantees, vows, contracts, pledges.

e.g.

I will drive the car according to British traffic laws.

Expressions

Apologies, thanks, welcomes, congratulations, condolences

e.g.

I offer my apologies for the party leadership due to health problems.

Declarations

Declarations, pronouncements

e.g.

Egyptian airways have decided to organize training courses on airplane piloting problems to its staff.

Appendix 8

This appendix presents the 7 lowest similarity pairs in Arabic short text (ASTSS-68) dataset

No.	Human Ratings	Standard Deviation	Short Text No. (Table 6.12)
1	0.01	0.06	5
2	0	0	26
3	0	0	46
4	0.02	0.13	53
5	0.02	0.13	55
6	0	0	60
7	0.02	0.13	61
Mean	**0.01**	**0.06**	

Appendix 9 presents the optimization dataset (ASTSS-21) used in the process of the optimization of parameters in the NasTa algorithms.

	Short Text Pairs	Human Ratings	ازواج الجمـــــل
1	I have fasted the three white days of each month in addition to the fasting in the month of Ramadan.	3.08	قمت بصيام الايام الثلاثة البيض من كل شهر اضافة الى صوم شهر رمضان
	I fasted for twenty days in addition to the days of fasting in the blessed month of Ramadan this year.		صُمْتُ عشرين يوماً إضافة إلى صيام أيام شهر رمضان المبارك في هذا العام.
2	The palm trees in Iraq will disappear entirely if the country does not care about the groves and keep what is left of them.	3.67	نخل العراق سيختفي كليا ان لم نهتم ببساتين البلاد ونحافظ على ما تبقى منها
	The tall palm trees in Iraq suffer severe negligence in the light of the deteriorating status of agriculture.		ان النخيل الباسقات في العراق تعاني اهمالا شديدا في ظل تدهور الواقع الزراعي
3	The devastating earthquakes that hit japan and East Asia are followed by high wave tsunami in the previous years and caused serious losses that made the whole world ready for them after every earthquake.	3.67	اعقبت الزلازل المدمرة التي ضربت اليابان وشرق آسيا الاعوام السابقة موجات تسونامي عالية جدا احدثت خسائر كبيرة جعلت العالم يتأهب لها بعد كل زلزال
	The tsunami is linked to the earthquake in East Asia and the people became afraid that each earthquake in this region will be followed by a new tsunami.		ارتبط التسونامي بالزلازل في شرق آسيا و أصبح الناس يخشون أن يتبع كل زلزال في هذه المنطقة تسونامي جديد
4	Did you know that the minister of education decided to give high school student free books?	3.33	هل تعلم أن وزير التعليم قرر أن يعطي طلاب الثانوية كتب بالمجان؟
	The government provides all students of primary and secondary schools with books and school supplies free for charge.		تجهز الحكومة جميع طلاب المدارس الابتدائية والثانوية بالكتب واللوازم المدرسية بالمجان

5	Venice is one of the most important and a beautiful Italian city that was erected on the sea and its unique location has attracted large number of tourists each year.	**3.69**	**البندقية من اهم واجمل المدن الايطالية التي اقيمت على مياه البحر ولموقعها المتميز هذا تجذب اعداد كبيرة من السياح كل عام**
	Venice is considered the oldest and most famous tourist cities in Italy, where it overlooks the sea and it is almost not devoid of visitors throughout the year.		**تعتبر البندقية من اشهر واقدم المدن السياحية في ايطاليا حيث تطل على البحر ولا تكاد تخلو من الزوار طوال السنة.**
6	O, Muslim prays standing, if you could not so pray while sitting down and if you could not then pray while laying on your side.	**3.33**	**ايها المسلم صل قائماً فإن لم تستطع فقاعداً فإن لم تستطع فعلى جنبك.**
	Originally, praying is to be done by a Muslim while standing and can be done while sitting for those people who have legitimate excuses		**الاصل في الصلاة ان يؤديها المسلم قائما ويصح الجلوس فيها لذوي الاعذار الشرعيه**
7	Take the friend a faithful brother and honest with you and will help you in a time of adversity.	**3.08**	**اتخذ الصديق اخا وفيا لك صادقا معك يعينك في وقت المحن والشدائد**
	I had many friends and did not like of them but the honest ones in their words and deeds.		**اتخذت اصدقاءاً كثر ولم يعجبني منهم الا الصادقون في اقوالهم وافعالهم**
8	Despite the passage of thousands of years, there are still traces of ancient civilizations based on our land up to this day.	**2.17**	**برغم مرور الاف الاعوام ما زالت آثار الحضارات القديمة قائمة على ارضنا الى يومنا هذا**
	The Arab land witnessed the dawn of civilization of humanity and there the first civilizations arose.		**شهدت الارض العربيه بزوغ الفجر الحضاري للانسانيه وعليها نشأت اولى الحضارات**
9	Financial and administrative corruptions are at the head of diseases that hinder the achievement of the desired development of our country.	**2.75**	**الفساد المالي والاداري على رأس الآفات التي تحول دون تحقيق التنمية المرجوه لبلدنا**
	There are a lot of problems plaguing the country, the first of which is the weakness of the government to achieve security.		**هناك الكثير من المشاكل التي تعصف بالبلاد وعلى رأسها ضعف الحكومة في تحقيق الامن**
10	Everyone who hears this announcement has to speed up to publicize the news through the media about the forced landing of	**2**	**على الجميع من يسمع هذا الإعلان المسارعة بنقل الخبر عبر وسائل الإعلام عن هبوط اضطراري لطائرة نقل ركاب بمطار جدة الدولي**

	the passenger plane in Jeddah International Airport.		
	The plane took off from Cairo, headed for Paris but due to the weather it was forced to land in Tunisia.		أقلعت الطائرة من القاهرة قاصدة باريس و لكن بسبب سوء الأحوال الجوية اضطرت للهبوط في تونس.
11	The head is located in the upper part of the human body and it contains the brain which is the control centre.	2.33	يقع الرأس في الجزء العلوي من جسم الإنسان ويحتوي على الدماغ الذي يعد مركز التحكم والسيطرة
	The head contains most of the senses enjoyed by human such as hearing, sight, smell and taste.		يحتوي الراس على معظم الحواس التي يتمتع بها الانسان كالسمع والبصر والشم والذوق
12	Most oriental women have large quantities of gold which they use for decoration and as a saving.	1.98	اغلب النساء الشرقيات يملكن كميات كبيرة من الذهب ويستخدمنه للزينة والتوفير.
	The rich celebrity artists have a lot of precious and beautiful jewellery.		تملك الفنانات الثريات ذوات الشهرة الواسعة الكثير من الجواهر الثمينة والجميلة
13	Sugar dissolves in water when the right amount of it is put while stirring continuously.	1.50	يذاب السكر في الماء عند وضع الكمية المناسبة مع التحريك بشكل مستمر
	Wait until coffee dissolves completely in warm water before adding the milk.		انتظري حتى تذوب القهوة تماما في الماء الدافئ قبل إضافة الحليب
14	My mother does not allow me to leave my room to play till she makes sure I have fully done my school homework.	1.25	لا تسمح والدتي لي بان اغادر غرفتي للعب حتى تتأكد من اني انهيت واجبي المدرسي تماما
	Make your bedroom first and then you can watch TV and use the computer.		رتب غرفة نومك اولا عندئذ يمكنك مشاهدة التلفاز واستخدام الحاسوب
15	My father is like the wedge; in my childhood he was a compassionate shepherd; in my youth he was a loyal friend and in his old age he was a wise mentor.	0	الوالد كالوتد فهو في طفولتي الراعي الحنون وفي شبابي الصديق الوفي وفي كبره الحكيم الناصح
	Yesterday, a huge fire occurred in the building of the institution facility and caused heavy losses.		وقع بالأمس حريق ضخم في مبنى المؤسسة تسبب في خسائر فادحة بالممتلكات
16	The pride of this nation is its youth’s commitment to the	0	فخر هذه الامة التزام شبابها بالاسلام وتطبيق تعاليمه في حياتهم العملية

	teaching of Islam and its application in their practical life.		
	The kite is one of the most popular games among children in summer.		تعتبر الطائرة الورقية من أكثر الألعاب الشعبية المنتشرة بين الأطفال في فصل الصيف
17	I promise before God to look after this boy and this girl until they reach the age of majority.	**0**	اتعهد امام الله بأن ارعى هذا الفتى وتلك الفتاة حتى يبلغا الرشد
	Do you think that the constant change in the climate allows the cultivation of palm trees in Britain in the future?		هل تظن أن التغير المستمر في المناخ سوف يسمح بزراعة النخل في بريطانيا في المستقبل؟
18	A rose gives its juice to insect and its perfume to all people and then withers away and dies silently.	**0**	تعطي الوردة رحيقها للحشرات وعطرها لجميع الناس ثم تذبل وتموت بصمت
	Migratory birds leave from south America to the north a distance of 200 thousands meters per week.		تقطع الطيور المهاجرة من جنوب امريكا الى شمالها مسافة 200 الف متر بالاسبوع
19	A smart writer employs the press in the service of his literature and does not leave it to devour his talent.	**0**	الاديب الذكي هو من يوظف الصحافة لخدمة ادبه لا ان يتركها تلتهم موهبته
	I bought a new car at great price after I got tired of purchasing old cars where there are many breakdowns.		اشتريت سيارة جديدة بثمن كبير بعد أن تعبت من شراء السيارت القديمة التي تكثر فيها الأعطال
20	I reward my son with coloured stickers when he compete his friend to encourage him in the sport.	**0**	أكافئ ابني بالملصقات الملونة عندما يسابق أصدقاءه لأشجعه على الرياضة
	Do you think that envier is loved by people or is he a castaway in the community?		هل تعتقد ان الحسود يحبه الناس ام انه منبوذ في المجتمع؟
21	I will go to him praising his generosity as a prelude to get back my money that he borrowed from me a year ago or more.	**0**	ساذهب اليه مادحا كرمه تمهيدا لأخذ مالي الذي استدانه مني قبل عام او اكثر
	People benefit from the whale oil which strengthens the bones and increases the vitality of the body.		ينتفع الناس بزيت الحوت الذي يقوي العظام ويزيد من حيوية الجسم

Appendix 10: Author Publications

ALMARSOOMI, F. A., O'SHEA, J. D., BANDAR, Z. A. & CROCKETT, K. A. Arabic word semantic similarity. Proceedings of World Academy of Science, Engineering and Technology, 2012. World Academy of Science, Engineering and Technology.

ALMARSOOMI, F. A., OSHEA, J. D., BANDAR, Z. & CROCKETT, K. AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity. Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, 2013. IEEE, 504-509.

Arabic Word Semantic Similarity

Faaza A, Almarsoomi, James D, O'Shea, Zuhair A, Bandar and Keeley A, Crockett

***Abstract*—** This paper is concerned with the production of an Arabic word semantic similarity benchmark dataset. It is the first of its kind for Arabic which was particularly developed to assess the accuracy of word semantic similarity measurements. Semantic similarity is an essential component to numerous applications in fields such as natural language processing, artificial intelligence, linguistics, and psychology. Most of the reported work has been done for English. To the best of our knowledge, there is no word similarity measure developed specifically for Arabic. In this paper, an Arabic benchmark dataset of 70 word pairs is presented. New methods and best possible available techniques have been used in this study to produce the Arabic dataset. This includes selecting and creating materials, collecting human ratings from a representative sample of participants, and calculating the overall ratings. This dataset will make a substantial contribution to future work in the field of Arabic WSS and hopefully it will be considered as a reference basis from which to evaluate and compare different methodologies in the field.

***Keywords*—** Arabic categories, benchmark dataset, semantic similarity, word pair, stimulus Arabic words.

INTRODUCTION

WORD semantic similarity (WSS) has grown to be an important part of natural language processing and information retrieval (IR) for many years. Semantic similarity is an essential component of numerous applications in the fields of artificial intelligence, psychology and computational linguistics, both in the academic community and industry. Examples comprise word sense disambiguation [1], IR [2], semantic search (to find pictures, documents, jobs and videos) [3], [4] and also in the seeking of biological macromolecules such as proteins and DNA [5].

Recently new measures have been proposed to calculate the semantic similarity between two short texts (STSS) of sentence length which rely largely on computing the similarity between words in both sentences [6]. These measures are promising techniques which can play a crucial role in the development of large number of applications. For example, in web page retrieval, STSS measure is used to improve retrieval effectiveness through the calculation of the similarities of page titles [7]. Text mining can also benefit from the use of STSS measure as a criterion to detect unseen knowledge from textual databases [8]. In the conversational agent / dialogue system, the employment of the STSS measure can greatly reduce the scripting process through the use of natural sentences instead of structural patterns of sentences [9].

F. Almarsoomi, J.D. O'Shea, Z. Bandar, and K. Crockett are with the Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M1 5GD, UK.

(e-mail: faaza-abdul.j.al-marsoomi@stu.mmu.ac.uk)

(e-mail: {j.d.oshea, z.bandar, k.crockett}@mmu.ac.uk).

These applications show that the calculation of semantic similarity between two words is a fundamental task which is frequently represented by similarity between concepts associated with the compared words.

There are a number of WSS measures [10] in the literature which have been evaluated through the use of the word similarity benchmark dataset before they are integrated into the complete system. Consistency of a WSS measure with human similarity ratings is employed to determine the quality of such measures. This is measured as the product-moment correlation coefficient computed between the set of human similarity ratings and those from the word similarity measure using a benchmark dataset [11].

To date, most of the reported word similarity measures are for English. However, there is no work done specifically for the Arabic language. Consequently, there is no Arabic word semantic similarity dataset. In order to improve the accuracy of a large number of Arabic applications [12], [13], it is important first to create an Arabic word semantic similarity dataset using the best possible available methods which will make a substantial contribution to future work in the field of Arabic WSS.

The focus of this paper is the production of the first word similarity benchmark dataset for Modern Standard Arabic (MSA) which is the formal language of the Arab world. Arabic is a Semitic language which is spoken by over 330 million people [14]. The Arabic alphabet uses 25 consonants and 3 long vowels which are written from right to left. These letters take different shapes based on their location in the word. Diacritics are written above or below the letters to represent the desired sound and to give a word the desired meaning [15]. Also Arabic words exhibit a complex internal structure, where words often incorporate affixes that mark grammatical inflections and clitics to signify different parts of speech [15].

In this paper, the first Arabic word similarity dataset is created which consists of 70 Arabic word pairs with human ratings. The methodology comprises of four fundamental steps which includes materials be gathered (word pairs), human ratings collected, overall ratings computed and the dataset validated. This methodology is described and illustrated in this paper.

The remaining sections of this paper are organized as follows: section 2 reviews the prior work on word semantic similarity measures and datasets. Section 3 describes the procedure of the production of the Arabic dataset which includes constructing the set of Arabic word pairs experiment and collecting human ratings experiment. Section 4 discusses the experimental results and compares the Arabic dataset with related work.

PRIOR WORK

A number of algorithms have been developed for measuring WSS; most of these measures are for the English language. The following sections provide a brief review of existing WSS measurements and the datasets used for comparing and evaluating them.

Word Semantic Similarity Measure

Existing WSS measures can be generally categorized into three groups based on the information source they exploit: Dictionary / Ontology based methods [16], [17] typically use the semantic information derived from knowledge bases to compute the WSS. Corpus-based methods [18] principally use the frequency of a word's occurrence to calculate WSS using statistical information derived from the large corpora. Hybrid methods [10], [19] calculate the WSS by combining multiple information sources. A detailed review of WSS measures can be obtained in [20], [21].

Word Similarity Benchmark Dataset

WSS measures have been evaluated using the word similarity benchmark dataset before they are integrated into the complete system. Two word benchmark datasets are commonly used for evaluating and comparing new developments, both of them for English language.

Rubenstein & Goodenough R&G [22] created the most influential word benchmark dataset for English. The procedure of the production of this dataset comprised of two steps. The first step involved generating 65 word pairs ranging from maximum to minimum similarity of meaning. A list of 48 English nouns represented in two columns (A and B) was employed to produce the 65 word pairs by selecting one word from column A and one from column B. The second step involved collecting the human similarity ratings of the 65 word pairs. 51 undergraduate participants were asked to assess the similarity between the word pairs based on how similar they were in meaning. The words pairs were ranked using a rating scale which ran from 0 (minimum similarity) to 4 (maximum similarity). However R&G dataset was published without justification for the specific choices of 48 nouns and the method of the combination of word pairs.

Miller & Charles (M&C) [23] replicated the R&G experiment and considered only 30 word pairs from the 65 word pairs of the R&G dataset to avoid an inherent bias towards low similarity. 38 undergraduate students (all Native English speakers) were asked to rank the 30 word pairs using a rating scale from 0 to 4. This experiment was performed 25 years after the R&G experiment, however the correlation between human ratings in the two datasets obtained a high value of 0.97. The M&C experiment was replicated by Resink [11] in 1995. The subset of 30 word pairs was ranked by the sample of 10 computer science graduate students and post-docs. This experiment obtained a high value correlation of 0.96 with M&C dataset. The results of these experiments show that the R&G dataset has indicated stability over the years. This stability illustrates that the use of human ratings could be a reliable reference for the purpose of comparison with computational methods.

The R&G dataset is still valuable 45 years after it was produced [21]. Therefore the R&G methodology is used as a general framework to produce the first word benchmark dataset for Arabic.

PRODUCTION OF THE ARABIC WORD SIMILARITY BENCHMARK DATASET

The methodology of the production of the Arabic dataset involved conducting two experiments. The aim of experiment 1 was to construct the set of Arabic word pairs, whilst the aim of experiment 2 was to collect the human similarity ratings. Furthermore, five fundamental hurdles were taken into consideration as a part of the Arabic word dataset design process:

1) Selecting a sample of participants representing the general human population. Because the dataset was created for Arabic, it was decided to use a representative sample of participants from different Arabic countries which signify the general population taking into account the subject knowledge, gender, and age.
2) Representation of the Arabic language with a delimited number of word pairs. A new method (described in section III.*A*) was used to select the stimulus Arabic words. These words were selected and presented in a way that contributes to the control of the range of semantic similarity (maximum to minimum) covered by the set of produced word pairs.
3) Selecting a representative sample of Arabic word pairs. This was achieved by conducting an experiment to generate the set of Arabic word pairs using human judgments.
4) Selecting the measurement scale. The type of statistical methods that can be applied to the similarity measures is defined based on the measurement scale used when they created. A ratio scale was used as a measurement scale in the prior work for both WSS measures and word similarity dataset [11], [22], and [23]. This dataset is intended to assess the accuracy of the algorithms (WSS) running on the scale from 0 (minimum similarity) to maximum which is a kind of ratio scale.
5) Collection of the ratings that precisely signify human conception of similarity. A combination of card sorting and semantic anchors (described in section III.*C*) was used as the most suitable procedure to collect human similarity ratings. This combination was selected based upon four experiments [24] which examined the impact of varying two factors, Order (randomize the order of the word pairs) and Anchors, on human ratings. The experimental results showed that one of the combinations, known as Card Sorting with Semantic Anchors was superior as it obtained significantly lower noise and a higher correlation coefficient.

A. *Selecting the Set of Stimulus Arabic Words*

The first step of the production of the Arabic dataset was to create a list of Arabic words which was presented later to produce the set of Arabic word pairs using human judgments. The decision was made to use categories known as category norms to select stimulus words for producing a list of Arabic words.

A category norm is defined as a set of words within the same theme, listed by frequency, which is created as responses by human participants to a specific category [25]. These categories consist of a large number of different themes used in many studies. For example, English category norms consist of 56 to 70 different themes used in 1600 projects after they were produced [26]. It was decided to employ category norms for selecting the set of stimulus words based on the two important features of these categories (a large number of different themes and a list of words within the same theme).

Due to the lack of category norms for the Arabic language, 27 Arabic categories were created and employed to select the stimulus Arabic words. As in category norms, the Arabic categories have different themes and consist of ordinary Arabic words. The words in each category are more similar to each other than to the words of other categories. The following steps illustrate the production of Arabic categories:

Step1. 22 categories were created to have the same themes as R&G to take advantage of four decades of experience with this dataset. The list of English words in the R&G experiment contains 48 nouns (24 pairs) for 22 different themes. This list was employed to create the 22 Arabic categories consisting of 22 different themes as follows:

1) For each English pair, the two nouns were translated into Arabic using the first meaning from an established English–Arabic dictionary [27]. To ensure translation accuracy, the translated nouns were checked by a professional translator and a lecturer fluent in both languages.
2) Based on the definition of two selected nouns [28], the Arabic category was given a specific name and a set of Arabic nouns (described in one word) within the same category theme were added for the production of the entire category.

For example, the English nouns (Gem and Jewel) were selected (same theme) and both were translated into (جوهرة) in Arabic. The Arabic category was created and called the Gemstones category (احجار كريمة) based on the definitions of jewel (*a precious stone used to decorate valuable things that you wear, such as rings or necklaces*) and gem (*a jewel or stone that is used in jewelry*). A set of Arabic words within the same category theme (Diamond /ماس, Pearl /لؤلؤ, Crystal /بلور, ...) were added to produce an entire category.

Some English nouns were omitted and not added to Arabic categories due to translation problems. First, some English nouns translated into the same Arabic word such as (*Gem and Jewel*) both translated as جوهرة in Arabic. Also some English nouns were translated into two Arabic words such as the English noun *Madhouse* in Arabic translates as مستشفى المجانين. Consequently, all translated nouns (described in two words or having the same translated word) were omitted and not added to the Arabic categories. Table I illustrates the English nouns and the reasons of omission.

As a result, 22 Arabic categories were produced from 48 translated nouns as shown in Table II.

Step2. 5 new categories were created to expand the 22 categories' themes and incorporate particular Arabic themes as shown in Table II. For example, the Arabic categories created in the first step have the type of male life stages category, to expand this theme and include male and female, the type of female life stages category was created. Religious events and type of lifestyle categories were produced to incorporate particular Arabic themes.

Using the Arabic categories created in step 1 and 2, the first two nouns were selected from each category to generate the set of 56 stimulus Arabic words which consisted of 27 different themes as shown in Table III.

TABLE I
ENGLISH NOUNS WITH THE REASONS OF OMISSION

	English Nouns	Arabic Nouns	The reason of omitting
1	Madhouse	مستشفى المجانين	Described in two words
2	Asylum	مستشفى المختلين	Described in two words
3	Gem / Jewel	جوهـره	Same translated word
4	Sage / Oracle	حكيـم	Same translated word
5	Slave / Serf	عبـد	Same translated word
6	Tool / Implement	اداة	Same translated word
7	Hill / Mound	تـل	Same translated word
8	Car / Automobile	سيـارة	Same translated word
9	Cock / Rooster	ديـك	Same translated word
10	Graveyard/ Cemetery	مقبـرة	Same translated word

TABLE II
THE LIST OF ARABIC CATEGORIES

	Categories Names	اسماء الفئات العربية
1	Medical Places	مواقع طبية
2	Handwritten text	نص مكتوب يدويا
3	Type of male's life stages	مراحل حياة الذكر
4	Member of the clergy	رجل دين
5	Transportation vehicles	مركبات نقل
6	Coastal area	منطقة ساحلية
7	Bird	طير
8	Type of furnishings	نوع من المفروشات
9	Source of a human body energy	مصدر طاقة جسم الانسان
10	Appliance for cooking	جهاز طهي
11	Gemstones	أحجار كريمة
12	Drinking utensil	ادوات او آنية للشرب
13	Geographic	جغرافية الارض
14	Parts of day	اجزاء اليوم
15	Type of equipment	نوع من معدات/ تجهيزات
16	Type of departure	نوع من رحيل/ مغادرة
17	Somebody practices witchcraft	شخص يمارس السحر
18	Wise person	شخص حكيم
19	Facial expressions	تعابير الوجهة
20	Material for tying things	مادة لربط الاشياء
21	Person in slavery	شخص في العبودية
22	Burial place	اماكن لدفن
23	Religious events	الاموات
24	Type of lifestyle	احداث دينية
25	Type of female life stages	نوع من نمط / اسلوب
26	Vacation activities	الحياة
27	Family members	مراحل حياة الانثى
		انشطة العطلات
		أعضاء العائلة

B. Experiment 1: Construction of the Set of Arabic Word Pairs

1. Participants

A sample of 22 Arabic native speakers was chosen to perform the task of generating the set of Arabic word pairs. The participants were from different Arabic countries which include: Iraq, Saudi Arabia, Jordan, Libya, and Palestine. The sample consisted of 10 academics (University lecturers) and 12 non-academics. They were 13 Science/Engineering vs. 9 Art/Humanities backgrounds. The average age was 34 years and the standard deviation (SD) was 6.3 with 13 female and 9 male.

2. Materials

A list of Arabic nouns was created through the use of the set of stimulus Arabic words (selected in section III.*A*). This was done by representing the set of 56 stimulus words in two columns (A and B) with each column containing 28 different Arabic words. As shown in Table III the list of Arabic nouns consists of 28 pairs of nouns and the nouns of each pair within the same theme such as *Hospital* and *Infirmary* (one noun (*Hospital*) in column A and one (*Infirmary*) in column B). The order of Arabic nouns in column B was randomized to minimize ordering effects. This list was presented to 22 Arabic participants to generate the set of Arabic word pairs ranging from high to low similarity of meaning.

Two recording sheets were used by 22 Arabic participants containing instructions (described in section *B.3*) to create two lists of Arabic word pairs which included: a High Similarity of Meaning list (HSM) containing 28 word pairs between strongly related and identical in meaning. A Medium Similarity of Meaning list (MSM) containing 32 word pairs between vaguely similar and very much alike in meaning while a low similarity of meaning list was selected randomly.

Because the list of Arabic nouns has 28 noun pairs (each pair has the same theme), the participants were requested to write 28 high similarity word pairs. Unlike the high and low similarity word pairs, it is relatively difficult for humans to write medium similarity word pairs. So, to increase the opportunity of obtaining medium similarity word pairs, the participants were asked to write 32 word pairs for (MSM) list.

3. Procedure

The list of Arabic nouns was employed to produce the set of Arabic word pairs by selecting one word from column A and one from column B based on the amount of similarity of meaning.

The participants were instructed to perform the following task.

1) Using the list of Arabic nouns, write a list of 28 Arabic word pairs that have HSM.
2) The Arabic word pairs always contain one word from column A and one from column B.
3) The HSM list contains word pairs between strongly related and identical in meaning.
4) Please write 28 word pairs because all uncompleted questionnaires must be ignored.

Following the same procedure, the participants were requested to write a list of 32 Arabic word pairs for MSM. Some notes were included in the instruction sheet which stated: "You can select any word from column A more than once with different words from column B to create new word pairs"; and also "Please do not write the same word pair more than once in the same sheet or between different sheets".

TABLE III
THE LIST OF ARABIC NOUNS

	Column A			Column B	
1	Hospital	مستشفى	1	Bus	باص
2	Signature	توقيع	2	Pigeon	حمامة
3	Boy	صبي	3	Grave	قبر
4	Master	سيد	4	Woodland	أحراش
5	Coach	حافلة	5	Vegetable	خضار
6	Coast	ساحل	6	Mountain	جبل
7	Hen	دجاجة	7	Means (noun)	وسيلة
8	Cushion	مسند	8	Diamond	الماس
9	Food	طعام	9	Travel (noun)	سفر
10	Stove	موقد	10	Lad	فتى
11	Gem	جوهرة	11	Infirmary	مشفى
12	Glass	كأس	12	Magician	مشعوذ
13	Forest	غابة	13	Midday	ظهيرة
14	Hill	تل	14	Sheikh	شيخ
15	Noon	ظهر	15	Pillow	مخدة
16	Tool	اداة	16	Thinker	مفكر
17	Journey	رحلة	17	Odalisque	جارية
18	Wizard	ساحر	18	Shore	شاطئ
19	Sage	حكيم	19	Endorsement	تصديق
20	Smile	ابتسامة	20	Laugh	ضحك
21	Cord	حبل	21	Oven	فرن
22	Slave	عبد	22	String	خيط
23	Sepulcher	ضريح	23	Tumbler	قدح
24	Feast	عيد	24	Young woman	شابة
25	Countryside	ريف	25	Walk (noun)	مشي
26	Run (noun)	جري	26	Sister	أخت
27	Brother	أخ	27	Fasting	صيام
28	Girl	فتاة	28	village	قرية

4. Experimental Results

A set of 70 Arabic word pairs were selected using the two lists of word pairs (HSM and MSM lists) generated through experiment 1 plus the list of low similarity word pairs which were selected randomly. Table IV illustrates the final set of Arabic word pairs, where the first and last columns represent the set of Arabic word pairs in English and Arabic. The second column contains the number of participants who chose the word pair.

1) The first 24 word pairs in table IV represent the high similarity word pairs which were selected using HSM list. Those word pairs were chosen by all the 22 participants.
2) The word pairs from 25 to 47 (23 pairs) represent the medium similarity word pairs which were chosen by more than half the participants.
3) The last 23 word pairs were selected to represent the low similarity word pairs. A combination of medium similarity candidate word pairs rated low by participants plus randomly selected low similarity word pairs (using the list of Arabic nouns) to allow for word pairs that were not chosen by the participants.

For each noun in the list of Arabic nouns, the frequency of appearance of this noun in the final set of Arabic word

pairs was calculated. The nouns which have an occurrence of more than two times were removed from the list of Arabic nouns to avoid a biased set of nouns from being used. The remaining Arabic nouns were used to generate a list of Arabic word pairs randomly. High and medium similarity word pairs already found by participants were removed. The remaining pairs were selected at random as they were good candidates for low similarity.

Experiment 2: Collection the Human Similarity Ratings

1. Participants

60 participants from different Arabic countries were asked to rank the set of 70 Arabic word pairs collected in Experiment 1. All were Arabic native speakers who had not taken part in Experiment 1 and they were from 7 Arabic countries which included: Iraq, Saudi Arabia, Egypt, Jordan, Kuwait, Libya, and Palestine.

The participants were equally balanced between students and non-students which they were39 Science/Engineering vs. 21 Art/Humanities backgrounds. The average age was 29 years and the standard deviation (SD) was 7.2 with an equal balance of male and female.

2. Materials

The set of 70 Arabic word pairs collected in experiment 1 were presented to Arabic participants to collect judgments on how similar they are in meaning. Each of 70 word pairs was printed on a separate card. Each participant was given an envelope containing 70 cards (the order of 70 cards was initially randomized to minimize the ordering effects) and 3 sheets which included: instructions for collecting the human rating, a similarity rating recording sheet and a personal information sheet.

TABLE IV
THE FINAL SET OF ARABIC WORD PAIRS

	Word Pairs		Participants	أزواج الكلمـــات	
1	Boy	Lad	22	فتـى	صبـي
2	Coast	Shore	22	شاطئ	ساحـل
3	Cushion	Pillow	22	مخـدة	مسنـد
4	Gem	Diamond	22	الماس	جوهرة
5	Glass	Tumbler	22	قـدح	كـأس
6	Forest	Woodland	22	أحراش	غابـة
7	Noon	Midday	22	ظهيـرة	ظهـر
8	Tool	Means	22	وسيلـة	اداة
9	Journey	Travel	22	سفـر	رحلـة
10	Smile	Laugh	22	ضحـك	إبتسامة
11	Countryside	Village	22	قريـة	ريـف
12	Girl	Young woman	22	شابـة	فتـاة
13	Signature	Endorsement	22	تصديق	توقيـع
14	Coach	Bus	22	بـاص	حافلـة
15	Hen	Pigeon	22	حمامة	دجاجة
16	Sepulcher	Grave	22	قبـر	ضريـح
17	Run	Walk	22	مشي	جـري
18	Hospital	Infirmary	22	مشفى	مستشفى
19	Master	Sheikh	22	شيـخ	سيـد
20	Wizard	Magician	22	مشعوذ	ساحـر
21	Feast	Fasting	22	صيـام	عيـد
22	Food	Vegetable	22	خضـار	طعـام
23	Stove	Oven	22	فـرن	موقـد
24	Hill	Mountain	22	جبل	تـل
25	Sage	Thinker	21	مفكر	حكيم
26	Cord	String	21	خيـط	حبـل
27	Slave	Odalisque	21	جاريـة	عبـد
28	Brother	Sister	21	أخت	أخ
29	Hen	Oven	20	فـرن	دجاجـة
30	Coach	Means	19	وسيلـة	حافلـة
31	Sage	Sheikh	18	شيـخ	حكيم
32	Girl	Sister	16	أخت	فتـاة
33	Journey	Shore	15	شاطئ	رحلـة
34	Coast	Mountain	14	جبل	ساحل
35	Master	Thinker	14	مفكر	سيـد
36	Coach	Travel	14	سفر	حافلـة
37	Food	Oven	14	فرن	طعام
38	Brother	Lad	13	فتـى	أخ
39	Girl	Odalisque	13	جاريـة	فتـاة
40	Slave	Lad	13	فتى	عبـد
41	Feast	Laugh	13	ضحك	عيـد
42	Hospital	Grave	12	قبر	مستشفى
43	Hill	Woodland	12	أحراش	تـل
44	Journey	Bus	12	باص	رحلـة
45	Tool	Tumbler	12	قـدح	اداة
46	Run	Shore	11	شاطئ	جري
47	Tool	Pillow	11	مخـدة	اداة
48	Sepulcher	Sheikh	10	شيخ	ضريـح
49	Cord	Mountain	9	جبل	حبـل
50	Gem	Young woman	8	شابة	جوهرة
51	Countryside	Vegetable	7	خضار	ريـف
52	Glass	Fasting	6	صيـام	كـأس
53	Forest	Shore	5	شاطئ	غابـة
54	Noon	Fasting	4	صيـام	ظهـر
55	Glass	Diamond	3	الماس	كـأس
56	Signature	String	2	خيط	توقيـع
57	Boy	Midday	1	ظهيرة	صبـي
58	Wizard	Infirmary	0	مشفى	ساحـر
59	Cushion	Diamond	0	الماس	مسنـد
60	Noon	String	0	خيـط	ظهـر
61	Boy	Endorsement	0	تصديـق	صبـي
62	Gem	Pillow	0	مخـدة	جوهرة
63	Cord	Midday	0	ظهيرة	حبـل
64	Countryside	Laugh	0	ضحك	ريـف
65	Hill	Pigeon	0	حمامة	تـل
66	Slave	Vegetable	0	خضـار	عبـد
67	Smile	Village	0	قريـة	إبتسامة
68	Stove	Walk	0	مشي	موقـد
69	Coast	Endorsement	0	تصديـق	ساحـل
70	Smile	Pigeon	0	حمامة	إبتسامة

3. Procedure

A combination of card sorting (sorting the cards based on the amount of similarity of meaning) and semantic anchors were used in this experiment to collect human judgments. A semantic anchor permits the participants to map a scale descriptor to each of the major scale points [24]. 5 semantic anchors for the 5 point rating scale listed in Table V were used in this experiment. The participants were requested to rate each word pair based on how similar they were in meaning after sorting the cards. Also they ranked each word pair using the 5 points rating scales which ran from 0.0 (unrelated in meaning) to 4.0 (identical in meaning).

The participants were asked to perform the following task:

1) Please sort the 70 cards into four groups according to the similarity of meaning. The HSM group contains word pairs between strongly related and identical in meaning, the two MSM groups contain word pairs vaguely similar or very much alike in meaning and low similarity contains word pairs unrelated in meaning.
2) The number of cards in each group is based on your judgment of each card.
3) Please check the cards in each group carefully; you may change a word pair from group to another at this stage.
4) Please rate each word pair according to the similarity of meaning using the rating scale points.

Furthermore, some notes were included in the instruction sheet which stated: "Please do not write values greater than 4.0 or less than 0.0. Also, you may rate more than one pair with the same value." And: "You can use the first decimal place to assign an accurate degree of similarity (for instance, if you think the similarity of word pair is between 2 and 3 you can assign a value such as 2.5)".

TABLE V
SEMANTIC ANCHORS

Rating scale	Semantic Anchors	
0	The word pairs are unrelated in meaning	زوج الكلمات لا يوجد ارتباط بينها في المعنى
1	The word pairs are vaguely similar in meaning.	زوج الكلمات بينها تشابه ضمني في المعنى
2	The word pairs are very much alike in meaning.	زوج الكلمات التي بينها تشابه واضح (اكثر من ضمني)
3	The word pairs are strongly related in meaning	زوج الكلمات التي بينها علاقة قوية في المعنى
4	The word pairs are identical in meaning	زوج الكلمات المترادفة او المتطابقة في المعنى

4. Experimental Results

Table VI contains the result of experiment 2 which represents the set of Arabic word pairs with a human similarity rating. The first and last pairs of columns represent the set of Arabic word pairs in English and Arabic. The third column contains the average of similarity rating collected from 60 Arabic native speakers.

DISCUSSION

The Benchmark Dataset

The human similarity ratings collected in experiment 2 are calculated as the mean of the judgments provided by the 60 participants for each of the Arabic word pairs as shown in Table VI.

The correlation coefficient is considered as a suitable measure for consistency [24]. The consistency between the set of human ratings and those obtained from the WSS algorithms is determined using the Pearson product-moment correlation coefficient which is considered suitable for measures created on a ratio scale [24]. Fig. 1 shows the correlation coefficients of 60 participants, where the consistency of similarity rating for each participant with the rest of group was determined using the Pearson product moment correlation coefficient. This was calculated by the leave-one-out resampling technique [11] for the ratings of each participant with all of the rest of the group.

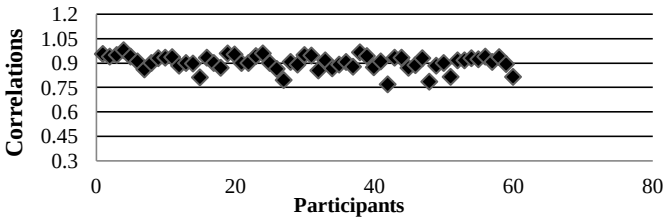


Fig. 1 Correlation coefficients of 60 participants

The average of the correlations of all participants on the Arabic dataset was calculated; this can be used to assess the performance of a computational (WSS) attempt to carry out the same task. Any WSS measure which equals or exceeds the average of the correlations of all participants is considered to be performing well. As shown in Table VII, the average of the correlations of all participants for the Arabic dataset is 0.902. The worst performing participant of 0.767 is considered as the lower bound for the expected performance whereas any machine measure coming close to the best performing participant at 0.974 would be considered as performing very well.

TABLE VII
PEARSON CORRELATION COEFFICIENT WITH MEAN HUMAN JUDGMENTS

	Correlation r
Average of the correlation of all participants	0.902
Best participant	0.974
Worst participant	0.767

Both high similarity and low similarity word pairs are subject to very consistent human judgments, as shown in Fig. 2 and Fig. 3. Unlike the low and high similarity word pairs, the human ratings of the medium similarity word pairs spread more evenly across the similarity range (0 to 4). Consequently, the medium similarity word pairs have higher values of SD than the other word pairs.

TABLE VI
THE SET OF ARABIC WORD PAIRS WITH HUMAN RATINGS

	Word Pairs		Human Ratings	أزواج الكلمـات	
1	Coast	Endorsement	0.03	تصديق	ساحـل
2	Noon	String	0.03	خيط	ظهر
3	Cushion	Diamond	0.06	الماس	مسند
4	Gem	Pillow	0.07	مخدة	جوهرة
5	Stove	Walk	0.07	مشي	موقـد
6	Cord	Midday	0.08	ظهيرة	حبل
7	Signature	String	0.08	خيط	توقيع
8	Boy	Endorsement	0.12	تصديق	صبـي
9	Boy	Midday	0.16	ظهيرة	صبـي
10	Slave	Vegetable	0.16	خضار	عبـد
11	Smile	Village	0.18	قريـة	إبتسامة
12	Smile	Pigeon	0.20	حمامة	إبتسامة
13	Wizard	Infirmary	0.22	مشفى	ساحر
14	Noon	Fasting	0.29	صيـام	ظهر
15	Hill	Pigeon	0.33	حمامة	تل
16	Countryside	Laugh	0.34	ضحك	ريف
17	Glass	Diamond	0.36	الماس	كـأس
18	Glass	Fasting	0.38	صيـام	كـأس
19	Cord	Mountain	0.54	جبـل	حبل
20	Hospital	Grave	0.83	قبر	مستشفى
21	Forest	Shore	0.86	شاطئ	غابـة
22	Gem	Young woman	0.87	شابة	جوهرة
23	Sepulcher	Sheikh	0.89	شيخ	ضريـح
24	Tool	Pillow	0.99	مخدة	اداة
25	Coast	Mountain	1.06	جبل	ساحل
26	Run	Shore	1.13	شاطئ	جري
27	Hill	Woodland	1.19	أحراش	تـل
28	Countryside	Vegetable	1.24	خضار	ريـف
29	Tool	Tumbler	1.32	قدح	اداة
30	Master	Thinker	1.36	مفكر	سيد
31	Feast	Laugh	1.36	ضحك	عيد
32	Hen	Oven	1.44	فرن	دجاجـة
33	Journey	Shore	1.47	شاطئ	رحلـة
34	Coach	Travel	1.60	سفر	حافلـة
35	Food	Oven	1.76	فرن	طعام
36	Slave	Lad	1.77	فتى	عبـد
37	Journey	Bus	1.83	باص	رحلـة
38	Girl	Odalisque	1.96	جاريـة	فتـاة
39	Feast	Fasting	1.96	صيـام	عيد
40	Coach	Means	2.07	وسيلـة	حافلـة
41	Brother	Lad	2.15	فتـى	أخ
42	Sage	Sheikh	2.26	شيخ	حكيم
43	Girl	Sister	2.38	أخت	فتـاة
44	Hill	Mountain	2.60	جبل	تـل
45	Hen	Pigeon	2.61	حمامة	دجاجة
46	Master	Sheikh	2.66	شيخ	سيد
47	Food	Vegetable	2.78	خضار	طعام
48	Slave	Odalisque	2.84	جاريـة	عبـد
49	Run	Walk	3.01	مشي	جـري
50	Brother	Sister	3.08	أخت	أخ
51	Cord	String	3.09	خيـط	حبل
52	Forest	Woodland	3.14	أحراش	غابـة
53	Sage	Thinker	3.30	مفكر	حكيم
54	Gem	Diamond	3.38	الماس	جوهرة
55	Cushion	Pillow	3.38	مخدة	مسند
56	Journey	Travel	3.39	سفر	رحلـة
57	Countryside	Village	3.41	قريـة	ريـف
58	Smile	Laugh	3.48	ضحـك	إبتسامة
59	Stove	Oven	3.55	فـرن	موقـد
60	Coast	Shore	3.56	شاطئ	ساحـل
61	Signature	Endorsement	3.58	تصديق	توقيـع
62	Tool	Means	3.68	وسيلـة	اداة
63	Noon	Midday	3.70	ظهيرة	ظهـر
64	Boy	Lad	3.71	فتـى	صبـي
65	Girl	Young woman	3.74	شابـة	فتـاة
66	Sepulcher	Grave	3.75	قبـر	ضريـح
67	Wizard	Magician	3.76	مشعوذ	ساحـر
68	Coach	Bus	3.80	بـاص	حافلـة
69	Glass	Tumbler	3.82	قدح	كـأس
70	Hospital	Infirmary	3.91	مشفى	مستشفى

For example, the word pair 46 (سيد شيخ) has SD 1.07 and the mean of human ratings 2.66. The distribution of the human ratings for this word pair should be grouped around a peak 2.66. In fact the module class is 3 and the distribution is relatively flat as shown in Fig. 4.

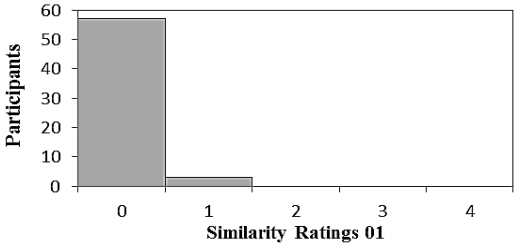


Fig. 2 Histogram of similarity ratings for word pair 01, SD=0.14.

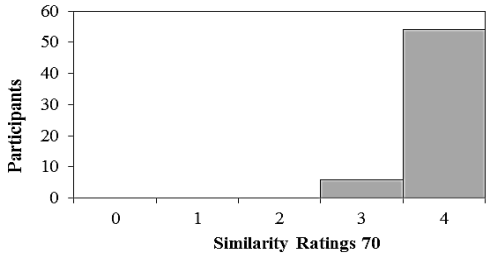


Fig. 3 Histogram of similarity ratings for word pair 70, SD=0.28.

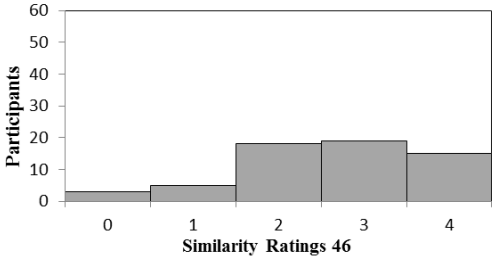


Fig. 4 Histogram of similarity ratings for word pair 46, SD=1.07.

A Comparison with the R&G Dataset

The most influential word dataset for English to R&G was used as a general framework for the production of the Arabic word dataset. In this section, a comparison is conducted between the two datasets to illustrate the differences between them.

1. Method of Selection of Materials

48 nouns (22 themes) to the R&G dataset were employed to make up the set of 65 word pairs in a variety of combinations which covered a range of semantic similarity values from high to low. However, the R&G dataset was published without justification for the specific choices of 48 nouns and the method of the combination of word pairs. The R&G dataset is skewed towards low similarity word pairs [23].

For this study 56 stimulus Arabic words (27 themes) were carefully selected through the use of 27 Arabic categories to generate the set of 70 Arabic word pairs. Semantic similarity judgments are an issue of human perception. Experiment 1 was used to create 70 word pairs spanning the similarity range based on human judgments to counter the bias towards low similarity in the R&G dataset.

2. Sampling the Population of Participants

The sample of participants used in the R&G experiment to collect human ratings was two groups of college undergraduates for a total of 51 participants. No information was provided on the composition of age or gender for each group and whether the sample of participants used in this experiment contained only native English speakers.

The sample of human population used in the Arabic dataset experiments is more representative than the R&G experiment. The value of a sample of participants selected to carry out a specific experiment could be reduced as a representative sample if there is a high homogeneity of participants and they are distant from the general population [24]. Consequently, the sample of Arabic participants was selected as a general population (students and non-students) from different Arabic countries taking account of the gender, age, and academic background factors. The sample was selected to balance gender (males and females), student and non-student, academic background (science/engineering vs. arts/humanities) and age to avoid a bias towards any element of these factors.

3. The Procedure of Collection Human Ratings

A card sorting technique was used for collecting human ratings in the R&G experiment. The 65 word pairs were presented to collect the human judgments. Each word pair was printed on a separate slip and the order of 65 slips was randomized before presentation. The participants were asked to sort the slips into order of similarity of meaning and each word pair was rated by assigning a value from 4.0- 0.0: the greater the similarity of meaning the higher the number.

A combination of card sorting with semantic anchors was used to collect human ratings in the Arabic dataset experiment, which is considered as the best currently known experimental practice. Each word pair in the dataset was printed on a separate card and the order of 70 cards was randomized before presentation. The participants were asked to sort the cards into four groups based on the similarity of meaning. The word pairs in each group were rated using a point rating scale (the points described by the semantic anchors) which ran from 0 (low similarity) to 4 (high similarity).

CONCLUSION

This paper has described the production of the first Arabic benchmark dataset for WSS algorithms. Though it is not possible to cover the language comprehensively in this dataset (70 word pairs), a new method was used to select the 56 stimulus Arabic words through the creation of 27 Arabic categories with 27 different themes to promote the best possible semantic representation. Unlike the prior work [22], participants were chosen to produce 70 word pairs which covered a range of word semantic similarity values from high (e.g. مستشفى - مشفى) to low (e.g. ساحل - تصديق). Human ratings were collected using the best currently known experimental practice and the statistical methods applied to calculate the overall ratings and defined the lower and upper bound for performance were the mean of human judgments and the Pearson Product-Moment correlation coefficient respectively. The sample of participants used in the Arabic dataset experiments were selected to get a balance and representation of the human population well beyond that of prior work. Furthermore, the procedure used for production of this dataset can be used by other Arabic researchers to extend the Arabic WSS benchmark dataset. Unfortunately, there are no WSS measures for Arabic, however the developments in English clearly point out the need for them. Also Arabic researchers are introducing the components required in terms of ontologies and corpora to produce such measures. Therefore, we present this dataset for future development and hopefully this will motivate Arabic researchers to start experimenting with Arabic word semantic similarity dataset. We are currently developing an Arabic word semantic similarity measure for calculating the similarity between concepts associated with the compared words in the Arabic lexical database known as Arabic wordnet [29]. The accuracy of this measure will be assessed using the Arabic word dataset developed in this paper.

REFERENCES

[1] S. Ravi, and M. Rada, "Unsupervised graph-based word sense disambiguation using measures of word semantic similarity," *In Proceedings of ICSC*, 2007.

[2] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. E. Milios, "Information retrieval by semantic similarity," *International Journal on Semantic Web and Information Systems*, vol. 2, no. 3, pp. 55-73, 2006.

[3] J. Davies, U. Krohn, and R. Weeks, "QuizRDF: Search technology for the semantic web," WWW2002 *workshop on RDF and Semantic Web Applications, 11th International WWW Conference WWW2002*, Hawaii, USA, 2002.

[4] Y. Aytar, M. Shah, and L. Jiebo, "Utilizing semantic word similarity measures for video retrieval," *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR08)*, pp. 1–8, Jun. 2008.

[5] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Measuring semantic similarity between Gene ontology terms," *Data & Knowledge Engineering*, vol. 61, no. 1, pp. 137-152, 2007.

[6] H. Chukfong, A. Masrah, M. Azmi, A. Rabiah and C. Shyamala, "Word sense disambiguation based sentence similarity," *Coling 2010: Poster Volume*, pp. 418–426, Beijing, Aug. 2010.
[7] E.K. Park, D.Y. Ra, and M.G. Jang, "Techniques for improving web retrieval effectiveness," *Information Processing and Management*, vol. 41, no. 5, pp. 1207–1223, 2005.
[8] J. Atkinson-Abutridy, C. Mellish, and S. Aitken, "Combining information extraction with genetic algorithms for text mining," *IEEE Intelligent Systems*, vol. 19, no. 3, 2004.
[9] K. O'Shea, Z. Bandar, and K. Crockett, "A Conversational agent framework using semantic analysis," *International Journal of Intelligent Computing Research (IJICR)*, vol. 1, no. 1, Mar. 2010.
[10] V. S. Zuber, and B. Faltings, "OSS: A semantic similarity function based on hierarchical ontologies," *In Proceedings of IJCAI*, pp. 551-556, 2007.
[11] P. Resnik, "Information content to evaluate semantic similarity in a taxonomy," *In Proceedings of IJCAI*, pp. 448-453, 1995.
[12] M. Diab, M. Alkhalifa, S. ElKateb, C. Fellbaum, A. Mansouri, and M. Palmer, "Semeval-2007 task 18: Arabic semantic labelling," *In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, 2007.
[13] M. Hijjawi, *ArabChat : an Arabic Conversational Agent*. PhD. Thesis, Department of Computing and Mathematics, Faculty of Science and Engineering, Manchester Metropolitan University, UK, 2011.
[14] A. Farghaly, K. Shaalan, "Arabic natural language processing: challenges and solutions," *ACMTransactions on Asian Language Information Processing*, vol. 8, no. 4, Article 14, 2009.
[15] N. Y. Habash, *Introduction to Arabic Natural Language Processing*. Graeme Hirst 2010. Morgan &Claypool, 2010, PP 11-12 & 39-41.
[16] M. Jarmasz, and S. Szpakowicz, "Roget's Thesaurus and semantic similarity," *In proceedings of the international conference on Recent Advances in Natural Language processing*, Borovetz, Bulgaria, pp. 212-219, 2003.
[17] R. Rada, H. Mili, M. Bicknell, and E. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 19, pp 17-30, 1989.
[18] D. Lin, "An Information-theoretic definition of similarity," *In Proceedings of Conference on Machine Learning*, pp. 296-304, 1998.
[19] Y. Li, Z. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871-882, 2003.
[20] T. Pedersen, V. S. Pakhomov, S. Patwardhan, and C.G. Chute, "Measures of semantic similarity and relatedness in the Biomedical Domain," *Journal of Biomedical Informatics*, vol. 40, PP. 288-299, 2007.
[21] G. Pirro, "Semantic similarity metric combining features and intrinsic information content," *Data & Knowledge Engineering*, vol. 68. pp. 1289-1308, 2009.
[22] H. Rubenstein, and J. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, Vol. 8, pp.627–633, 1965.
[23] G.A. Miller, and W.G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, pp.1–28, 1991.
[24] J.D. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "Benchmarking Short Text Semantic Similarity," *Int. J. Intelligent Information and Database Systems*, vol. 4, no. 2, pp. 103-120, 2010.
[25] W.F. Battig, and W.E. Montague, "Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms," *Journal of Experimental Psychology Monographs*, vol. 80, PP. 1–46, 1969.
[26] J.P. Van Overschelde, K.A. Rawson, and J. Dunlosky (2004), "Category norms: An updated and expanded version of the Battig and Montague (1969) norms," *Journal of Memory and Language*, vol. 50, pp. 289–335, 2004.
[27] B. Munir, *AL-MAWRID: A Modern English-Arabic Dictionary*. Dar EL-ILMILMALAYIN, Beirut, Lebanon. Edition 11, 1977. www.malayin.com.
[28] J. Sinclair, *Collins Cobuild English Dictionary for Advanced Learners*, 3rd edn. Harper Collins, New York, 2001.
[29] S. Elkateb, W. Black, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum, "Building a WordNet for Arabic," *In Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity

Faaza A. Almarsoomi, James D. O'Shea, Member, IEEE, Zuhair Bandar, and Keeley Crockett, Senior Member, IEEE

School of Computing, Mathematics and Digital Technology
Manchester Metropolitan University
Manchester, United Kingdom
faaza-abdul.j.al-marsoomi@stu.mmu.ac.uk

***Abstract*— Semantic similarity is an essential component of numerous applications in fields such as natural language processing, artificial intelligence, linguistics, and psychology. Most of the reported work has been done in English. To the best of our knowledge, there is no word similarity measure developed specifically for Arabic. This paper presents a method to measure the semantic similarity between two Arabic words in the Arabic knowledge base. The semantic similarity is calculated through the combination of the common and different attributes between the Arabic words in the hierarchy semantic net. We use a previously developed Arabic word benchmark dataset to optimize and evaluate the Arabic measure. Experimental evaluation indicates that the Arabic measure is performing well. It has achieved a correlation value of 0.894 compared with the average value of human participants of 0.893 on evaluation dataset.**

Keywords-semantic similarity; arabic language; benchmark dataset; dialogue systems

I. INTRODUCTION

To date, no prior work has been reported on word semantic similarity for the Arabic language. This paper presents a novel algorithm for measuring the semantic similarity of Arabic word pairs. The only way to evaluate such measure meaningfully is by comparison with human perception. Consequently, this work uses a dataset of human ratings published in [1] during the early stages of developing the algorithm.

The ability to formalize and quantify the intuitive notion of semantic similarity between words is a problem with a long history in artificial intelligence, computational linguistics and psychology [2]. The difficulty lies in how to obtain an effective method to emulate the human judgment process of word semantic similarity through processing and combining several information sources.

Semantic similarity is vital for numerous applications in many research fields. Examples comprise word sense disambiguation [3], Information Retrieval [4], and semantic search to find pictures, documents, and jobs [5] [6]. Word semantic similarity has also been proposed as component for measuring the similarity between two short texts of sentence length, which can play a crucial role in the development of the performance of the bulk of applications relying on it [7].

Assessing semantic similarity between words is frequently represented by similarity between concepts associated with the compared words. Interest in automatic word semantic similarity started in the 1960s particularly for the English language. Since then, a number of algorithms have been proposed using a variety of approaches, which can generally be viewed in terms of the information source they exploit: Corpus-based methods principally use the frequency of a word's occurrence to calculate the similarity between words using statistical information derived from the large corpora. Knowledge based methods typically use the semantic information derived from knowledge bases to assess the similarity between a pair of words. The work has been extended to other European languages and is beginning in Thai.

The technique used in this paper makes use of the semantic knowledge base known as Arabic WordNet (AWN) [8]. Firstly, it extracts common and different attributes of the concepts associated with the compared words in the taxonomy of AWN. Secondly, it calculates the similarity based on the relationship between common and different attributes of the compared words extracted in the first step.

The second contribution of the work in this study is the optimization of parameters in the algorithm through partitioning the Arabic dataset [1] into training and evaluation sets, which is a known problem in English [9].

Consistency of the AWSS measure with human ratings is employed to identify its quality. The possible indicative value and bounds of performance expected from the AWSS measure have been calculated as the average, worst and best performances of human participants on the Arabic evaluation dataset.

In section II, we review some prior works briefly. An AWSS model for calculating similarity between Arabic words is presented in section III. Section IV describes the process of the production of Arabic word benchmark datasets. The experimental results are discussed in section V and the paper is concluded with proposing some future works in section VI.

II. PRIOR WORK AND BACKGROUND

The two important factors in creating an AWSS measure are what can be drawn from prior works in the English language and the availability of Arabic linguistic resources for use in an algorithm.

978-1-4799-0652-9/13/$31.00 © 2013 IEEE

A. Prior Work for English

As mentioned in the introduction, no prior work has been reported on AWSS measure. However, related work on English word similarity measures provides a starting point.

Rada et al. [10] utilizes the minimum path length connecting the concepts containing the compared words as a measure for calculating the similarity of words. Their work is considered the basis of edge counting-based methods.

Resnik's measure [2] is the first to combine an ontology and a corpus together. Some modifications have been performed to improve the pure information content measure of the original work of Resnik. Jiang and Conrath [11] presented a hybrid method on the basis of the edge-based notion through adding the information content as a decision factor. The same elements of Jiang and Conrath method are used by Lin [12] to calculate semantic similarity but in a different fashion. Lin proposed a new formula derived from information theory, which combines information content of the compared words and assuming their independence.

Leacock and Chodorow [13] proposed a method for measuring the similarity based on the shortest path length between two concepts using IS-A link, taking into consideration the maximum depth of the noun taxonomy.

Hirst and St-Onge [14] proposed a measure that considers the two concepts are semantically close if a path that is not too long and that does not modify its direction too often connects their synsets in WordNet. The semantic relatedness measure sets different weights for different links in the semantic knowledge base in order to generate a model more closely to human performance.

Li et al. [9] presented different strategies to calculate the semantic similarity using multiple information sources, which are the shortest path length, depth and local density. The best strategy obtained the best result that combines the shortest path and depth nonlinearly.

As introduced above, different approaches use different information sources and, thus, result in different performance levels. The commonly used information sources in the reported similarity measures are shortest path length between compared words, depth in the taxonomy hierarchy, information content, and semantic density of compared words. The first group of the proposed measures used the information sources directly as a metric of word similarity while the second used a particular information source without considering the contribution of others. A third group claimed that the information sources should be properly processed and combined. A knowledge based method [9] proposed based on the third notion obtained the best performance among the reported word similarity measures according to [4] [15], which they carried out a comparison between the performance of these measures.

B. Arabic Knoweledge Resources

Arabic is a Semitic language which is spoken and written by more than 300 million people in the world and is considered a highly derivational and inflexional language. However, little work has been done on developing linguistic resources for Arabic NLP, especially knowledge rich resources such as ontologies that can support Arabic semantic similarity. Furthermore, only theoretical models are presented and no implementation is available of these projects e.g. the work in [16]. This work describes an ontological representation for the Arabic Language. This ontology is relevant because its design is based on Semitic template root-based lexical principles, which represent the Arabic language features but no implementation is available.

AWN is the only free lexical resource for modern standard Arabic [8]. It is based on the design and contents of Princeton WordNet (PWN) [17] and can be mapped onto PWN as well as a number of other wordnets. The AWN structure consists of four principal structures. First, the items represent conceptual entities including synsets, ontology classes and instances. Second, a word entity represents a word sense. Third, a form entity contains lexical information. Fourth, a link connects in a relation two items. Moreover, the mapping of wordnet to the Suggested Upper Merged Ontology (SUMO) [18] provides opportunities to use the semantic side in some Arabic applications. The latest version of AWN consists of 11,270 synsets containing about 23,496 Arabic words.

Because of the prior work reviewed in sections *A* and *B*, this study utilizes a knowledge-based method to identify the score of similarity between two Arabic words using the latest version of AWN.

III. Arabic Semantic Similarity Model

According to [19], the similarity between two concepts is identified by humans through comparing their common and different attributes. These attributes are considered for simulating the process of human judgments. Therefore, the similarity between two Arabic words is calculated based on the relationship between different and common attributes of compared words in the semantic knowledge base. The semantic knowledge base such as AWN is constructed in a lexical hierarchy where words are connected with concepts by well-defined types of relations. The concepts at lower levels of the lexical hierarchy have more concrete semantics and stronger similarity which can be used to identify the different attributes of compared words. This is done through the calculation of the shortest path between the concepts containing the compared words. The concepts at upper levels of the hierarchy possess more general semantics and less similarity between them. This intuition can be utilize to identify the common attributes of the compared words through the computation of the depth of the concept (Lowest Common Subsumer (LCS)) that subsume the concepts containing the two words.

Fig. 1 illustrates a portion of AWN noun hierarchy. The shortest path length between أب father and أم mother is **2** and the concept شخص parent is called LCS for the words أب father and أم mother; while the shortest path between جد grandparent and أب father is **6**. In this case, we could say the أم mother is more similar to أب father than جد grandparent is to أب father. Also in this figure, the shortest path length between جد grandparent and تاجر عملة money_handler is **5**, less than from جد grandparent to أب father, but we should not say جد grandparent is more

similar to تاجر عملة money_handler than to father. This case illustrates the importance of the depth of LCS where the similarity of compared words grows higher if the depth of LCS increases as we go deeper in a lexical hierarchy.

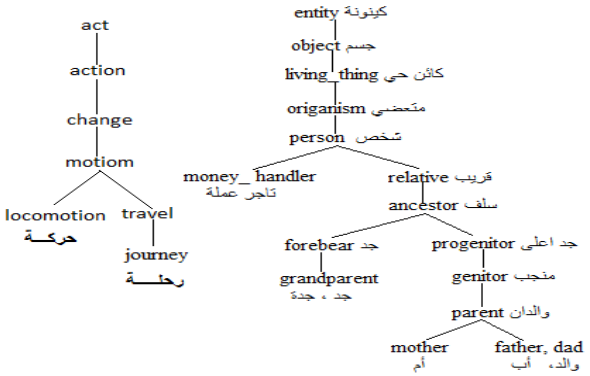


Figure 1. A portion of Arabic wordnet noun hierarchy.

In this paper, the semantic similarity is identified using information sources extracted from AWN, which are length and depth.

Given two words w1 and w2, the semantic similarity between them as in [9] can be defined as a function of the attributes of path length (different attributes) and depth (common attributes) as follows:

$$s(w1,w2) = F(f1(l), f2(d)) \quad (1)$$

Where, l is the length of the shortest path between w1 and w2. d is the depth of the LCS of w1 and w2 in a lexical hierarchy. $f1$ and $f2$ are transfer functions of path and depth respectively.

The similarity interval is [0, 1]. When l =0, the similarity of s (w1, w2) = 1 which implies that the similarity is inversely proportional to path length.

For example, in Fig. 1, أب father and والد dad are in the same concept and length between them is 0. This case implies that the two words have the same meaning. Therefore, $f1$ is set to be a monotonically decreasing function of l and is selected in exponential form to meet l constraints.

When d=0, there are no common attributes between the compared words and the similarity of s (w1, w2) = 0. As shown in Fig.1, رحلة journey and أب father are classified under separate substructure and no LCS subsumes the compared words and hence the similarity between them is **0**. Furthermore and as shown in the example of جد grandparent and تاجر عملة money-handler, the similarity grows higher if the depth of LCS of compared words increases in a lexical hierarchy. To meet this constraints, $f2$ is set to be increasing function of d.

In this paper, the overall similarity is calculated using the following nonlinear formula [9]:

$$sim(w1, w2) = e^{(-\alpha * l)} * \tanh(\beta * d) \quad (2)$$

Where, α and β are the length and depth factors respectively which signify the contribution of the length l and depth d. l can be calculated using (3):

$$l = d1+d2 - (2*d) \quad (3)$$

Where $d1$ and $d2$ are the depth of w1 and w2 respectively.

IV. Experiment

A. Production of the Data Set

The quality of a computational word similarity measure can be identified through the investigation of its performance compared with human common sense. This can be assessed by calculating word similarity on an Arabic word set with human judgments. The first Arabic word dataset produced by [1] is employed to assess the accuracy of Arabic word similarity measure. Creating this dataset required a substantial and sound experimental methodology which was partitioned into three major stages include creating a List of Arabic Words (LAW), constructing the set of Arabic word pairs and collecting the human ratings for pairs of words.

The major step of the production of the Arabic dataset is selecting a set of stimulus words that represents the Arabic language for evaluating the AWSS measures effectively. This was achieved by carefully selecting 56 stimulus words through the employment of categories known as category norms. Category norm is a set of words, listed by frequency and generated as responses by human participants to a specific theme [20]. Due to the lack of category norms for the Arabic language, 27 Arabic categories were produced to cover different semantic themes and contain ordinary Arabic words. These categories were employed to generate a set of 56 stimulus Arabic words by selecting the first two words from each category.

LAW was created through the use of the set of stimulus words. This was done by representing the 56 stimulus words into two columns each column contains a word from each theme.

One of the fundamental obstacles to the production of Arabic word dataset is selecting a sample of word pairs that precisely represents the huge range of word pairs that can be generated from LAW. This problem was solved by conducting an experiment to construct a representative sample of word pairs based on human judgments. LAW was presented to 22 Arabic Native speakers from 5 Arabic countries to construct a set of word pairs covering the range of similarity of meaning (high to low). The participants were asked to create two lists of word pairs which include high and medium similarity of meaning. The final set of Arabic word pairs contains 70 pairs of words which were selected using high and medium similarity word pairs lists generated by participants plus the low similarity word pairs list selected randomly.

The second experiment was conducted for collecting the human similarity ratings for the set of 70 word pairs collected in experiment 1. This experiment used a sample of 60 Arabic Native speakers from 7 Arabic countries who had not taken part in the first experiment. Each of 70 word pairs was printed on a separate card and those cards were presented to participants for rating how similar the word pair on each card was in meaning. The order of 70 cards was randomized before presentation. Each of 60 participants was requested to sort the 70 cards based on the

similarity of meaning and rate them using scales which ranged from 0.0 (low similarity) to 4.0 (high similarity).

Finally, each of the 70 Arabic word pairs was assigned a semantic similarity score calculated as the mean of the ratings provided by 60 Arabic native speakers. This dataset is the first of its kind for Arabic and should become a gold standard for evaluation and comparison of future AWSS measures. The set of Arabic word pairs with human ratings and the detailed procedure for creating this data set are published in [1].

B. Application of the Data Set

The evaluation process of AWSS measure requires identifying the optimal value of AWSS measure parameters. Therefore, the Arabic dataset has been divided into two sets one known as train dataset was employed to tune the AWSS measure parameters and another denoted as evaluate dataset was used to assess its accuracy. Each dataset consists of 35 word pairs spanning the similarity of meaning range from maximum to minimum, which were selected as follows.

1) The original Arabic dataset consists of 24 low similarity, 24 medium similarity and 22 high similarity word pairs. Therefore, each sub-dataset contains 12 low similarity, 12 medium similarity and 11 high similarity word pairs.
2) For each similarity class within the same sub-dataset, the word pairs were selected ranging the similarity of meaning from low to high.

Only 30 word pairs of each sub-datasets have been used in our experiments. The reason is that, Some Arabic words have not been added to the current version of AWN such as موقد stove, ساحر wizard, مستشفى hospital… etc. In addition, some Arabic words do not have complete senses such as the Arabic word ضحك laugh, which has just two senses in the current version of AWN. While the sense (laugh as a facial expression) has not been added to the current version.

The word pairs in the train and evaluate datasets are listed with human ratings in Table I and Table II, respectively. The bold word pairs have not been used in our experiments.

C. Tuning

The AWSS measure parameters (α and β) have been tuned using the training dataset to find the optimal values within the interval [0, 1]. Given the initial value of each parameter, the training dataset word pairs were run using the AWSS measure to produce the machine similarity ratings in the range 0 to 1. The correlation coefficient between the human ratings of Arabic dataset and those obtained from the AWSS measure was computed. The values of the Arabic measure parameters were changed to obtain a set of correlation coefficients. Then the parameters with the strongest correlation coefficient were considered as the optimal parameters. In our experiment, the strongest correlation coefficient was obtained at $\alpha = 0.162$ and $\beta = 0.234$.

The word pairs in evaluate dataset were run using the identified optimal parameters for producing the machine similarity ratings. The correlation coefficient was calculated again between the machine and human ratings for pairs of words in the evaluation dataset to assess the accuracy of the AWSS measure. Table II shows the human similarity ratings with the corresponding machine similarity ratings on evaluate dataset.

V. RESULTS AND DISCUSSION

The possible bounds of performance expected from an Arabic word measure have been calculated as the average, worst and best performances of human participants on the evaluate dataset as shown in Table III. This was done using the leave-one-out resampling technique [2] to calculate the correlation coefficient of each of 60 participants with the rest of group. The consistency of Arabic measure with human perception was identified by computing the correlation coefficient between the average rating of human participants and the machine ratings as shown in Table III.

TABLE I. TRAIN DATASET WORD PAIRS WITH HUMAN RATINGS

Word Pairs		Human Ratings	أزواج الكلمـــات	
Cushion	Diamond	0.01	الماس	مسند
Gem	Pillow	0.02	مخدة	جوهرة
Cord	Midday	0.02	ظهيرة	حبل
Signature	String	0.02	خيط	توقيع
Boy	Endorsement	0.03	تصديق	صبي
Boy	Midday	0.04	ظهيرة	صبي
Smile	Pigeon	0.05	حمامة	ابتسامة/بسمة
Noon	Fasting	0.07	صيام	ظهر
Countryside	Laugh	0.08	ضحك	ريف
Glass	Fasting	0.10	صيام	كأس
Hospital	**Grave**	**0.21**	**قبر**	**مستشفى**
Gem	Young woman	0.22	شابة	جوهرة
Run	Shore	0.28	شاطئ	جري
Hill	Woodland	0.30	أحراش	تـل
Countryside	Vegetable	0.31	خضار	ريف
Master	Thinker	0.34	مفكر	سيد
Feast	Laugh	0.34	ضحك	عيد
Hen	**Oven**	**0.36**	**فرن**	**دجاجة**
Slave	Lad	0.44	فتى	عبـد
Journey	Bus	0.46	باص	رحلة
Girl	Odalisque	0.49	جارية	فتاة
Brother	Lad	0.54	فتى	أخ
Sage	Sheikh	0.57	شيخ	حكيم
Hen	Pigeon	0.65	حمامة	دجاجة
Brother	Sister	0.77	أخت	أخ
Sage	Thinker	0.83	مفكر	حكيم
Gem	Diamond	0.85	الماس	جوهرة
Journey	Travel	0.85	سفر	رحلة
Smile	**Laugh**	**0.87**	**ضحك**	**ابتسامة/بسمة**
Stove	**Oven**	**0.89**	**فرن**	**موقد**
Signature	Endorsement	0.90	تصديق	توقيع
Noon	Midday	0.93	ظهيرة	ظهر
Girl	Young Woman	0.94	شابة	فتاة
Coach	Bus	0.95	باص	حافلة
Hospital	**Infirmary**	**0.98**	**مشفى**	**مستشفى**

TABLE II. EVALUATE DATASET WORD PAIRS WITH HUMAN AND MACHINE RATINGS

Word Pairs		Human Ratings	Machine Ratings	أزواج الكلمـات	
Coast	Endorsement	0.01	0.0	تصديق	ساحـل
Noon	String	0.01	0.27	خيط	ظهر
Stove	**Walk**	**0.02**	-	**مشي**	**موقـد**
Slave	Vegetable	0.04	0.06	خضار	عبد
Smile	Village	0.05	0.0	قريـة	ابتسامة/بسمة
Wizard	**Infirmary**	**0.06**	-	**مشفى**	**ساحـر**
Hill	Pigeon	0.08	0.06	حمامة	تـل
Glass	Diamond	0.09	0.05	الماس	كأس
Cord	Mountain	0.13	0.17	جبل	حبل
Forest	Shore	0.21	0.17	شاطئ	غابـة
sepulcher	Sheikh	0.22	0.06	شيخ	ضريح
Tool	Pillow	0.25	0.32	مخدة	أداة
Coast	Mountain	0.27	0.45	جبل	ساحل
Tool	Tumbler	0.33	0.54	قدح	أداة
Journey	Shore	0.37	0.0	شاطئ	رحلة
Coach	Travel	0.40	0.0	سفر	حافلة
Food	**Oven**	**0.44**	-	**فـرن**	**طعـام**
Feast	Fasting	0.49	0.17	صيام	عيد
Coach	Means	0.52	0.38	وسيلة	حافلة
Girl	Sister	0.60	0.37	اخت	فتـاة
Hill	**Mountain**	**0.65**	-	**جبـل**	**تـل**
Master	Sheikh	0.67	0.67	شيخ	سيـد
Food	Vegetable	0.69	0.53	خضار	طعـام
Slave	Odalisque	0.71	0.93	جاريـة	عبـد
Run	Walk	0.75	0.60	مشي	جـري
Cord	String	0.77	0.70	خيط	حبـل
Forest	Woodland	0.79	0.82	أحراش	غابـة
Cushion	Pillow	0.85	0.82	مخدة	مسند
Countryside	Village	0.85	0.82	قرية	ريف
Coast	Shore	0.89	0.89	شاطئ	ساحل
Tool	Means	0.92	0.93	وسيلة	أداة
Boy	Lad	0.93	0.95	فتى	صبي
Sepulcher	Grave	0.94	0.82	قبر	ضريح
Wizard	**Magician**	**0.94**	-	**مشعوذ**	**ساحـر**
Glass	Tumbler	0.95	0.89	قدح	كـأس

TABLE III. PERFORMANCE OF AWSS MEASURE ON EVALUATE DATASET

On Evaluate Data Set	Correlation r
Arabic word similarity measure	0.894
Average of the correlation of all participants	0.893
Best participants	0.970
Worst participants	0.716

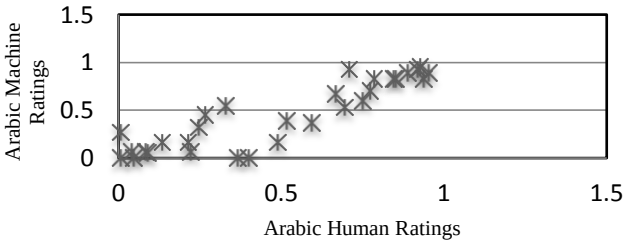


Figure 2. The correlation between human and machine ratings.

The AWSS measure obtained a good value of Pearson correlation coefficient (r = 0.894) with the human judgments as shown in Fig. 2. The AWSS measure is performing well at (r= 0.894) with the average value of the correlations of human participants (r = 0.893). Furthermore, the performance of the Arabic word measure is substantially better than the worst human (lower bound) performance at (r=0.716).

There are some anomalies that exist in the results. For example, the word pair ظهر خيط (Noon, String) was ranked a very low similarity value in human judgments, while a medium similarity rating was obtained by the AWSS measure. This is because; the Arabic word خيط string can be used in Arabic as evidence of the time especially with the dawn. The majority of participants chose this word as a string on account of comparing it with Noon, very little chose it as evidence of the time. In AWN, both (ظهر خيط) have a sense indicates to the time and the algorithm chose the time sense giving higher rating than the human rating.

In contrast, the word pair سفر حافلة (coach, travel) was given a human rating value higher than the machine similarity rating. An explanation is provided through looking at the noun hierarchy in AWN. A fragment of noun hierarchy is shown in Fig. 3 which involves all the senses of the word pair سفر حافلةAs can be observed, coach and travel are classified under separate substructures that mean no connection (no common features) between them in AWN hierarchy. This led to obtain a very low machine rating value. The substructure containing the word حافلة - coach has the synset (conveyance) as shown in Fig. 3; another sense for this word in AWN is transportation. The hyponym of transportation is (transportation – movement – change – action – act). It would be more sensible if the substructure including the word حافلة - coach were put under the class of transportation as shown in Fig 4. If so the synset (change) would connect between the word pair سفر حافلة (coach travel) and the machine similarity rating would have been closer to human assessment.

In consequence of the nature of AWN organization scheme, the structure of wordnet hierarchy may produce a bias towards a particular distance computation. This problem hopefully will be solved in future with the new versions of AWN. For the same reason the word pair رحلة شاطئ (journey shore) obtained machine rating lower than human similarity rating.

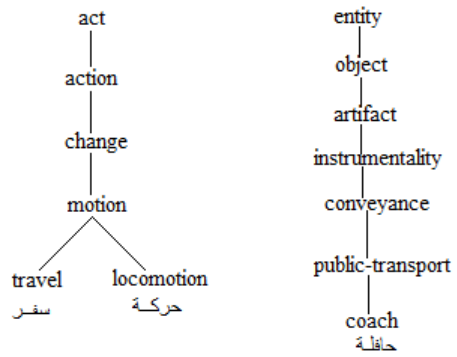


Figure 3. Fragment of the Arabic wordnet for the word pair travel and coach

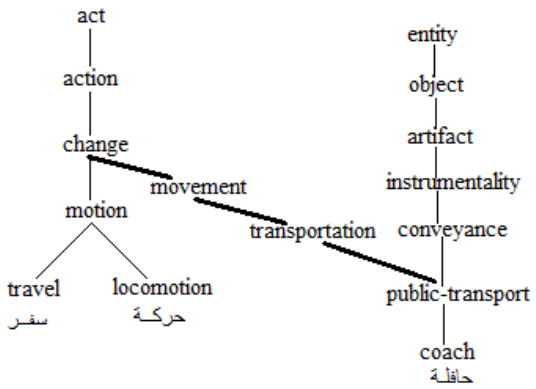


Figure 4. Another sense for the word conveyance

VI. CONCLUSION

There is no implementation available for an Arabic ontology that represents the Arabic language features as Semitic template root-based lexical principles. An Arabic word similarity measure was presented in this study for calculating the similarity between two Arabic words using ontology built based on models for English and Indo-European languages. The different and common attributes was extracted form Arabic ontology and combined to calculate the score of similarity between two Arabic words. The optimal value of Arabic measure parameter was identified using the training dataset produced in this work. An experiment was carried out on a dataset of word pairs with human similarity ratings. The correlation against the human similarity ratings on evaluating dataset is 0.894 compared with the human average of 0.893 for the same partition of the data. Analysis of the complete dataset [1] suggests that a correlation of 0.902 is a reasonable expectation. Despite using only half of the data, the approach has still scored substantially better than the worst human in [1].

In the future, we are planning to overcome the Arabic measure limitations through the benefit from the mapping of AWN to SUMO [18], which provides additional knowledge that may help to improve the similarity score. In addition, we would like to create a short text semantic similarity measure for modern standard Arabic using AWSS measure created in this study.

REFERENCES

F. Almarsoomi, J. O'Shea, Z. Bandar, and K. Crockett, "Arabic word semantic similarity," in *Proc. ICALLL (WASET)*, Dubai, UAE, 2012, vol. 70, pp. 87–95.

P. Resnik, "Using information content to evaluate semantic similarity," in *the 14th Int. Joint Conf. Artificial Intelligence*, Montreal, 1995, pp. 448–453.

S. Ravi, and M. Rada, "Unsupervised graph-based word sense disambiguation using measures of word semantic similarity," in *Proc. ICSC*, 2007, pp. 363–369.

A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. E. Milios, "Information retrieval by semantic similarity," *Int. J. Semantic Web and Information System*, vol. 2, no. 3, pp. 55–73, 2006.

J. Davies, U. Krohn, and R. Weeks, "Quizrdf: search technology for the semantic web," in *the 37th HICSS workshop on RDF and Semantic Web Applications*, USA, 2004.

Y. Aytar, M. Shah, and L. Jiebo, "Utilizing semantic word similarity measures for video retrieval," in *IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR08)*, 2008, pp.1–8.

H. Chukfong, A. Masrah, M. Azmi, A. Rabiah and C. Shyamala, "Word sense disambiguation based sentence similarity," in *the 23rd Int. Conf. COLING Poster Volume*, Beijing , 2010, pp. 418–426.

S. Elkateb, W. Black, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum, "Building a wordnet for arabic," in *the 5th Int. Conf. on Language Resources and Evaluation (LRE)*, Italy, 2006.

Y. Li, Z. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans. Knowledge and Data Engneering*, vol. 15, no. 4, pp. 871-882, Aug. 2003.

R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 17–30, Jan. 1989.

J. Jiang and D. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. Int. Conf. on Research in Computational Linguistics (RCOLING)*, Taiwan, 1997, pp. 19–33.

D. Lin, "An information-theoretic definition of similarity," in *the 15th Int. Conf. on Machine Learning (ICML)*, Madison, 1998, pp. 296–304 .

C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," in *Proc. Fellbaum'98*, 1998, pp. 265–283.

G. Hirst and D. St-Onge, "Lexical chains as representations of context for the detection and correction of malapropisms," in *Proc. Fellbaum'98*, 1998, pp. 305–332.

G. Pirro, "Semantic similarity metric combining features and intrinsic information content," *Data & Knowledge Engineering*, vol. 68, no. 11, pp. 1289-1308, 2009.

F. Belkridem, and A. El Sebai, "An ontology based formalism for the arabic language using verbs and derivatives," *Communications of IBIMA*, vol. 11, pp. 44–52, 2009.

G.A. Miller, "WordNet: a lexical database for english," *Comm. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

I. Niles, and A. Pease, "Linking lexicons and ontologies: mapping wordnet to the suggested upper merged ontology," in *Proc. Int. Conf. on Information and Knowledge Engineering*, Las Vegas, Nevada, 2003.

X. Liu, Y. Zhou, and R. Zheng, "Measuring semantic similarity in wordnet," in *Proc. ICMLC*, 2007, pp. 123–128.

W. Battig, and W. Montague, "Category norms for verbal items in 56 categories: a replication and extension of the connecticut category norms," *J. Experimental Psychology Monographs*, vol. 80, pp. 1–46, 1969.